sensors

Machine Health Monitoring and Fault Diagnosis Techniques

Edited by
Shilong Sun, Changqing Shen and Dong Wang

Printed Edition of the Special Issue Published in *Sensors*

www.mdpi.com/journal/sensors

MDPI

Machine Health Monitoring and Fault Diagnosis Techniques

Machine Health Monitoring and Fault Diagnosis Techniques

Editors

Shilong Sun
Changqing Shen
Dong Wang

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Shilong Sun
Harbin Institute of Technology
China

Changqing Shen
Soochow University
China

Dong Wang
Shanghai Jiao Tong University
China

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Sensors* (ISSN 1424-8220) (available at: https://www.mdpi.com/journal/sensors/special_issues/MHMFDT).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-7332-8 (Hbk)
ISBN 978-3-0365-7333-5 (PDF)

Contents

About the Editors

Shilong Sun

Dr. Shilong Sun currently works as an Assistant Professor at the Harbin Institute of Technology, Shenzhen now. He received a Ph.D. degree from the Department of Systems Engineering & Engineering Management at the City University of Hong Kong in 2018. He nurtures keen interests in vibration energy harvesting design, fault diagnosis, and prognosis, decision-making with Artificial Intelligence, deep learning for industrial data. Now he focuses on the research areas of equipment remaining life estimation with deep learning and smart energy harvesting techniques, respectively.

Changqing Shen

Changqing Shen (Senior Member, IEEE) received B.S. and Ph.D. degrees in Instrument Science and Technology from the University of Science and Technology of China, Hefei, China, in 2009 and 2014, respectively, and a Ph.D. degree in Systems Engineering and Engineering Management from the City University of Hong Kong, Hong Kong, in 2014. He is currently a Professor at the School of Rail Transportation, Soochow University, Suzhou, China. His research interests include signal processing and machine-learning-based fault diagnosis. He is an Associate Editor for the IEEE Open Journal of Instrumentation and Measurement.

Dong Wang

Dong Wang received his Ph.D. degree from the City University of Hong Kong in Hong Kong, China, in 2015. He was a Senior Research Assistant, Postdoctoral Fellow, and Research Fellow at the City University of Hong Kong. He is currently an Associate Professor at the Department of Industrial Engineering and Management of Shanghai Jiao Tong University in Shanghai, China, where he is also involved at the State Key Laboratory of Mechanical System and Vibration. His research interests include sparse and complex measures, signal processing, prognostics and health management, condition monitoring and fault diagnosis, statistical learning, machine learning, statistical process control, and nondestructive testing. Dr. Wang is an Editorial Board Member for Mechanical Systems and Signal Processing and an Associate Editor for IEEE Transactions on Instrumentation and Measurement, Measurement, IEEE Sensors Journal, and Journal of Dynamics, Monitoring, and Diagnostics.

Preface to "Machine Health Monitoring and Fault Diagnosis Techniques"

Machine health monitoring and fault diagnosis are crucial aspects of modern industrial systems. The timely and accurate detection and diagnosis of faults in machines can greatly improve the efficiency and reliability of industrial processes, reduce downtime, and prevent catastrophic failures. In recent years, advances in sensing technologies, data analytics, and machine learning have driven significant research activity in machine health monitoring and fault diagnosis.

This Special Issue, "Machine Health Monitoring and Fault Diagnosis Techniques", presents the latest research and developments in the field from leading experts, covering a wide range of topics: signal processing, feature extraction, fault diagnosis algorithms, prognostics, and sensor data-level approaches. The papers in this Special Issue address the challenges of machine health monitoring and fault diagnosis in various application domains, including manufacturing, bearing, resistance welding process, and diesel engines.

The editors hope that this Special Issue will stimulate further research and development in the field. We express our gratitude to all authors who contributed to this Special Issue and to the reviewers who provided their valuable feedback and insights. We also thank the editorial staff and the publisher for their support in bringing this Special Issue to fruition.

We hope that readers will find this Special Issue informative and inspiring and that it will contribute to the advancement of machine health monitoring and fault diagnosis techniques.

Shilong Sun, Changqing Shen, and Dong Wang
Editors

sensors

MDPI

Editorial

Editorial for Special Issue: Machine Health Monitoring and Fault Diagnosis Techniques

Shilong Sun [1,2,*], Changqing Shen [3] and Dong Wang [4]

[1] School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen 518055, China
[2] Guangdong Provincial Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics, Shenzhen 518055, China
[3] School of Rail Transportation, Soochow University, Suzhou 215131, China
[4] The State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai 200240, China
* Correspondence: sunshilong@hit.edu.cn

Citation: Sun, S.; Shen, C.; Wang, D. Editorial for Special Issue: Machine Health Monitoring and Fault Diagnosis Techniques. *Sensors* **2023**, *23*, 3493. https://doi.org/10.3390/s23073493

Received: 13 March 2023
Revised: 20 March 2023
Accepted: 24 March 2023
Published: 27 March 2023

Machine health monitoring and fault diagnosis have played crucial roles in automatic and intelligent industrial plants. Machine-learning-, deep-learning-, and artificial-intelligence-based intelligent fault diagnoses are essential in industrial settings, in order to help reduce the downtime that is caused by machine failures. These techniques can be integrated with advanced sensor technologies to enhance the accuracy of their results. Additionally, some specific artificial intelligence algorithms can help to identify potential problems and alert engineers on time. However, there are still several issues that require further investigation, with intelligent fault diagnosis methodologies being among them, e.g., early fault detection features, the few-shot sample machine learning algorithm, data augmentation techniques for deep learning, the data fusion method for domain adaptation, feature representation with self-supervision, and interpretable deep learning algorithms. Ultimately, machine health monitoring and fault diagnosis techniques are essential tools for ensuring a machine's safety and efficiency.

This Special Issue aims to highlight the state-of-the-art techniques that are used for machine health monitoring and fault diagnosis, especially for intelligent fault diagnosis algorithm development, fault feature extraction, and intelligent machine monitoring.

This Special Issue has received 26 manuscripts, 18 of which have been accepted, and 8 of which were rejected by the peer-review processes. These accepted manuscripts can be divided into four types: (1) status detection; (2) degradation process; (3) fault diagnosis; and (4) failure detection with sensors. Their details have been illustrated as follows:

1. Status Detection

In [1], a novel automated algorithm for the modal parameter identification of rotating machinery was described. The innovation of this study was that it targeted a rotor mode and is applicable to different systems and environments. This algorithm extracted the rotor and fundamental frequency damping ratios from a stability diagram that was given a user-defined parameter.

In [2], a multimodal process monitoring method, which was based on variable-length sliding window-mean augmented Dickey–Fuller (VLSW-MADF) test and a dynamic locality-preserving principal component analysis (DLPPCA), was proposed.

The work that was developed in [3] presented a new algorithm for impeller blade monitoring, based on a relative shaft vibration signal measurement and analysis, which was designed to run from a long-term perspective as part of a remote monitoring system in order to automatically track a natural blade frequency and its amplitude.

In [4], a method using multidimensional k-means for the condition monitoring of electrode wear was established. With the aid of this method, the relationship between the

serial time data of the resistance and the mechanical properties variation of the electrodes was described.

2. Degradation Process

A method for evaluating bearing performance degradation, by using an adaptive sensitive feature selection and multi-strategy optimization support vector data description (SVDD), was developed in [5]. In combining the technique for order preference by similarity with an ideal solution (TOPSIS) and K-medoids, monotonicity, correlation, and robustness indicators were used to determine an adaptive sensitive feature set for evaluation.

In [6], a new cluster migration distance (CMD) algorithm was proposed to address the problem in which traditional performance degradation indicators cannot accurately describe degradation trending on time. By calculating the offset trajectory of a feature cluster centroid in a continuous bearing running process, the CMD can appropriately handle the complex and variable features in the fault evolution process of a water pump bearing.

In [7], they used a ring oscillator (RO)-based test structure to extract data and build a dataset that could be used to predict aging trends and determine the primary aging mechanisms of 28 nm FPGAs. Moreover, they proposed a method to correct the temperature-induced measurement errors that are found in accelerated tests. Furthermore, they employed four machine learning (ML) technologies that were based on accurate measurement datasets, in order to predict these FPGA aging trends.

In [8], a method based on an improved particle swarm optimization (PSO) was proposed to analyze the bearing performance degradation. This proposed method can effectively resolve the problems of online parameter selection and the low predictive accuracy of long–short time memory (LSTM) methods. A kernel joint approximate diagonalization of eigen-matrices (KJADE) method was used to fuse the bearing vibration signals and form an effective feature vector, and an SS was calculated to acquire a performance degradation index. Subsequently, an improved PSO algorithm was used to optimize the LSTM parameters, in order to obtain an optimal performance degradation prediction model.

3. Fault Diagnosis

In [9], a model for data augmentation was proposed. This study proposed a method for the unbalanced fault diagnosis of rotating machinery that combined time–frequency feature oversampling (TFFO) with a convolutional neural network (CNN). The proposed model built a balanced dataset by simultaneously expanding time domain signals and time–frequency domain features, and by performing a comprehensive data expansion from different dimensions.

The work in [10] proposed a rolling bearing fault diagnosis method that was based on the whale gray wolf optimization algorithm–variational mode decomposition–support vector machine (WGWOA-VMD-SVM), which was designed to solve the unclear fault characteristics of rolling bearing signals, owing to its nonlinear and nonstationary characteristics. A rolling bearing signal was decomposed using variational mode decomposition (VMD), and a support vector machine (SVM) was used as the fault diagnosis model.

In [11], a novel model was proposed for an intelligent bearing fault diagnosis in rotating machinery. The main contribution of this model is the construction of an effective image dataset using a combination of an improved fast kurtogram (IFK) that was based on nonlinear mode decomposition (NMD) and a gramian angular field (GAF). The proposed model used IFK to achieve a high computational efficiency and improve its SNR. Next, GAF provided images that preserved the absolute temporal relationships of the signals, so that the CNN could perform a fault classification.

In [12], a novel intelligent rolling bearing fault diagnosis method, based on a Markov transition field (MTF) and a residual network, was proposed. Encoding one-dimensional time series signals into two-dimensional images with a Markov transition field preserved the time dependence of the raw signals and discarded the prior knowledge, in order to set

the parameters during the conversion. On this basis, a residual network was applied to identify the fault types through image classification.

In [13], a rolling bearing fault diagnosis method was proposed based on successive variational mode decomposition (SVMD) and an energy concentration and position accuracy (EP) index. The EP index effectively indicated a target mode for the characteristic fault information, and a line-searching method that was guided by the EP index optimized the balancing parameter of the SVMD.

In [14], a method of compressing and reconstructing diesel engine vibration signals was proposed by using sparse Bayesian optimization block learning, which combined a compressive sensing technology with the fault diagnosis. Its specific steps were as follows. The method achieved the optimal compression and reconstruction efficiency, was verified by several assessment indicators, and had a good classification accuracy. However, there was still room for improvement, particularly in the signal repair and noise reduction preprocessing, as well as in the integration of the algorithm into the data acquisition hardware.

In [15], an integrated vision transformer (ViT) model, which was based on wavelet transform and a soft voting method for the bearing fault diagnosis, was proposed. A vibration signal was decomposed into sub-signals in different frequency bands using discrete wavelet transform (DWT), and was transformed into time–frequency representation (TFR) maps using continuous wavelet transform (CWT). Multiple individual ViT models were used to preliminarily diagnose the faults, and a final diagnosis result was obtained by a fusion method that was based on the soft voting method.

4. Failure Detection with Sensor

In [16], the authors analyzed the heat generation of normal bearings and faulty bearings during the operation, and the influence of different working conditions on the heat generation of these bearings. In this study, based on the structural characteristics of the bearings, a new transient temperature analysis model for damaged bearings was established, considering the influence of the thermal–solid coupling effect on the bearing structure.

In [17], the authors analyzed different types of sensor faults for the fault detection of a healthy drive, using a variety of index-based methods. In total, seven main indices were employed and analyzed for the sensor fault diagnosis, including the moving mean, average, root mean square, energy, variance, the first-order derivative, the second-order derivative, and an auto-correlation-based index.

In [18], a novel virtual sensor for predictive maintenance, which was called a miniterm, was introduced. One of its main advantages was that its installation did not involve a large financial outlay. The evolution of the TAV (technical availability), mean time to repair (MTTR), EM (number of work orders (emergency orders/line stop)), and OM (labor hours in EM) showed a very important improvement, as the number of mini-terms increased and the Miniterm 4.0 system became more reliable.

5. Conclusions

The theme of this Special Issue focuses on machine health monitoring and fault diagnosis techniques, especially intelligent fault diagnosis. This Special Issue highlights 18 articles that can be divided into four categories: condition monitoring [1–4], degradation process prediction [5–8], intelligent diagnostic algorithms [9–15], and sensor fault detection [16–18]. In addition to the traditional bearing vibration signals, the research objects include the electrode signals, blade vibration signals, diesel engine vibration signals, and bearing heat signals. Therefore, in the field of fault diagnosis, in addition to the traditional bearing vibration signal analysis, other objects or signals can also be used as diagnostic features, which is worth studying. Regarding the algorithm design, the development of artificial intelligence algorithms also provides new solutions for other signal analyses and processing. Artificial intelligence algorithms and multi-sensor signals, combined

with intelligent fault diagnosis algorithms, will be a very important development trend in the future.

Acknowledgments: We want to thank all authors for their valuable contributions to this Special Issue; without their efforts, it could not be so successful. All manuscripts published in this Special Issue have passed two rounds of peer-reviewing progresses, and a final approval decision was given after the second round. The papers selected for this Special Issue represent the quality, breadth, and depth of machine health monitoring and fault diagnosis techniques.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dreher, N.R.; Storti, G.C.; Machado, T.H. Automated Operational Modal Analysis for Rotating Machinery Based on Clustering Techniques. *Sensors* **2023**, *23*, 1665. [CrossRef] [PubMed]
2. Wang, S.; Wang, Y.; Tong, J.; Chang, Y. Fault Monitoring Based on the VLSW-MADF Test and DLPPCA for Multimodal Processes. *Sensors* **2023**, *23*, 987. [CrossRef] [PubMed]
3. Liska, J.; Vasicek, V.; Jakl, J. A Novel Method of Impeller Blade Monitoring Using Shaft Vibration Signal Processing. *Sensors* **2022**, *22*, 4932. [CrossRef] [PubMed]
4. Ibáñez, D.; Garcia, E.; Soret, J.; Martos, J. An Unsupervised Condition Monitoring System for Electrode Milling Problems in the Resistance Welding Process. *Sensors* **2022**, *22*, 4311. [CrossRef] [PubMed]
5. Feng, Z.; Wang, Z.; Liu, X.; Li, J. Rolling Bearing Performance Degradation Assessment with Adaptive Sensitive Feature Selection and Multistrategy Optimized SVDD. *Sensors* **2023**, *23*, 1110. [CrossRef] [PubMed]
6. Zhai, Z.; Zhu, Z.; Xu, Y.; Zhao, X.; Liu, F.; Feng, Z. Cluster Migration Distance for Performance Degradation Assessment of Water Pump Bearings. *Sensors* **2022**, *22*, 6809. [CrossRef] [PubMed]
7. Li, Z.; Huang, Z.; Wang, Q.; Wang, J.; Luo, N. Implementation of Aging Mechanism Analysis and Prediction for XILINX 7-Series FPGAs with a 28-nm Process. *Sensors* **2022**, *22*, 4439. [CrossRef] [PubMed]
8. Hu, Y.; Wei, R.; Yang, Y.; Li, X.; Huang, Z.; Liu, Y.; He, C.; Lu, H. Performance Degradation Prediction Using LSTM with Optimized Parameters. *Sensors* **2022**, *22*, 2407. [CrossRef] [PubMed]
9. Zhang, L.; Liu, Y.; Zhou, J.; Luo, M.; Pu, S.; Yang, X. An Imbalanced Fault Diagnosis Method Based on TFFO and CNN for Rotating Machinery. *Sensors* **2022**, *22*, 8749. [CrossRef] [PubMed]
10. Zhou, J.; Xiao, M.; Niu, Y.; Ji, G. Rolling Bearing Fault Diagnosis Based on WGWOA-VMD-SVM. *Sensors* **2022**, *22*, 6281. [CrossRef] [PubMed]
11. Lee, C.-Y.; Zhuo, G.-L.; Le, T.-A. A Robust Deep Neural Network for Rolling Element Fault Diagnosis under Various Operating and Noisy Conditions. *Sensors* **2022**, *22*, 4705. [CrossRef] [PubMed]
12. Yan, J.; Kan, J.; Luo, H. Rolling Bearing Fault Diagnosis Based on Markov Transition Field and Residual Network. *Sensors* **2022**, *22*, 3936. [CrossRef]
13. Guo, Y.; Yang, Y.; Jiang, S.; Jin, X.; Wei, Y. Rolling Bearing Fault Diagnosis Based on Successive Variational Mode Decomposition and the EP Index. *Sensors* **2022**, *22*, 3889. [CrossRef]
14. Bai, H.; Wen, L.; Ma, Y.; Jia, X. Compression Reconstruction and Fault Diagnosis of Diesel Engine Vibration Signal Based on Optimizing Block Sparse Bayesian Learning. *Sensors* **2022**, *22*, 3884. [CrossRef] [PubMed]
15. Tang, X.; Xu, Z.; Wang, Z. A Novel Fault Diagnosis Method of Rolling Bearing Based on Integrated Vision Transformer Model. *Sensors* **2022**, *22*, 3878. [CrossRef] [PubMed]
16. Sun, Y.; Zhang, C.; Zhao, X.; Liu, X.; Lu, C.; Fei, J. Transient Thermal Analysis Model of Damaged Bearing Considering Thermo-Solid Coupling Effect. *Sensors* **2022**, *22*, 8171. [CrossRef] [PubMed]
17. Narzary, D.; Veluvolu, K.C. Multiple Sensor Fault Detection Using Index-Based Method. *Sensors* **2022**, *22*, 7988. [CrossRef] [PubMed]
18. Garcia, E.; Montés, N.; Llopis, J.; Lacasa, A. Miniterm, a Novel Virtual Sensor for Predictive Maintenance for the Industry 4.0 Era. *Sensors* **2022**, *22*, 6222. [CrossRef] [PubMed]

sensors

MDPI

Article

Automated Operational Modal Analysis for Rotating Machinery Based on Clustering Techniques

Nathali Rolon Dreher, Gustavo Chaves Storti and Tiago Henrique Machado *

School of Mechanical Engineering, University of Campinas, Campinas 13083-970, Brazil
* Correspondence: tiagomh@fem.unicamp.br

Abstract: Many parameters can be used to express a machine's condition and to track its evolution through time, such as modal parameters extracted from vibration signals. Operational Modal Analysis (OMA), commonly used to extract modal parameters from systems under operating conditions, was successfully employed in many monitoring systems, but its application in rotating machinery is still in development due to the distinct characteristics of this system. To implement efficient monitoring systems based on OMA, it is essential to automatically extract the modal parameters, which several studies have proposed in the literature. However, these algorithms are usually developed to deal with structures that have different characteristics when compared to rotating machinery, and, therefore, work poorly or do not work with this kind of system. Thus, this paper proposes, and has as its main novelty in, a new automated algorithm to carry out modal parameter identification on rotating machinery through OMA. The proposed technique was applied in two different datasets to enable the evaluation of the robustness to different systems and test conditions. It is revealed that the proposed algorithm is suitable for the accurate extraction of frequencies and damping ratios from the stabilization diagram, for both the rotor and the foundation, and only one user defined parameter is required.

Keywords: automated operational modal analysis; rotating machinery; hydrodynamic bearings; rolling bearings; hierarchical clustering

Citation: Dreher, N.R.; Storti, G.C.; Machado, T.H. Automated Operational Modal Analysis for Rotating Machinery Based on Clustering Techniques. *Sensors* **2023**, *23*, 1665. https://doi.org/10.3390/s23031665

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 23 October 2022
Revised: 27 November 2022
Accepted: 2 December 2022
Published: 2 February 2023

1. Introduction

Structural Health Monitoring (SHM) is the process of implementing a damage identification strategy for aerospace, civil, and mechanical engineering infrastructures [1]. SHM strategies have been employed in recent decades in order to improve the infrastructure's lifetime and safety. According to Lynch, Farrar, and Michaels [2], SHM can be divided into damage detection, prognostic, and risk assessment. The first step usually consists of collecting the structure's response over extended periods of time, followed by a data normalization for signal processing purposes, extracting damage-sensitive features, and finally, implementing a robust method for damage detection using the extracted features.

The structure's modal parameters can be used as damage-sensitive features in damage detection since they are based on parameters that are modified in the presence of damage. The modal parameters can be extracted by modal testing, using either Experimental Modal Analysis (EMA) or Operational Modal Analysis (OMA). EMA extracts the modal parameters considering that both inputs and outputs are measured whereas OMA obtains these parameters only from the measured outputs of the system. Whereas EMA requires equipment to excite the system and needs to take the system out of operation, OMA's premise is that the environmental loads acting upon the system excite it with an approximate white noise signal and do not require the system to go out of operation. Since the idea of SHM involves the constant monitoring of the structure, EMA is more adequate to an initial study of the modal parameters and OMA becomes an alternative to the monitoring while in operation.

There are numerous techniques to apply OMA, and one of the most employed is the Stochastic Subspace Identification (SSI). This technique has some advantages when compared to the others, such as the presence of noise truncating mechanisms based on Singular Value Decomposition (SVD), the solution of the identification problem by means of linear algebra tools, which avoids non-linear optimization problems and results in a lower computational cost, and the possibility of using weighting matrices to improve the method's performance in the presence of noise or weakly excited modes [3]. Yet, the quality of the estimation relies on a correct choice of SSI parameters, e.g., the number of block rows, the weighting matrices, the system order, etc. Theoretically, the system order could be estimated inspecting the number of singular values that are different from zero, and the number of block rows could be determined through a direct relation between the system order and the number of outputs. However, this approach is not suitable for real data because noise is usually present in the measurements and because of the structures' complexity. Hence, different approaches were proposed to handle real data. In order to solve the number of block rows problem, Reynders and De Roeck [4] proposed a relationship between the sampling frequency and the lowest frequency of interest, and the so-called stabilization diagram was proposed to deal with the fact that the system order is unknown and to better visualize and interpretate the technique's results.

To build this diagram, the SSI method is performed with increasing system orders and the obtained poles are plotted on a diagram of frequency vs. system order. In other words, it is possible to identify modes that stabilize in frequency, damping ratio and mode shape with increasing orders, which usually represent physical modes of the system. Because of that, the stabilization diagram is inevitably composed of groups of the physical modes that can be identified by the group of machine learning techniques called clustering. Given that the system order can be overestimated, several spurious and mathematical poles also appear on the diagram because of the model's attempt to better fit the experimental data, making it harder to identify the system's modal parameters. Moreover, when the system is excited by periodic signals, as rotating machinery in operation, several poles appear on the diagram at the fundamental and harmonic frequencies of the signal, and the alignment of these poles can be misinterpreted as modes of the system. Considering this, a procedure that combines clustering techniques with other signal processing techniques to automatically interpretates the stabilization diagram is essential and can promote a better and more reliable parameter extraction and would allow the SHM to be carried out without much user interaction in the selection of the system's modes.

Magalhães, Cunha, and Caetano [5] proposed an algorithm for the automatic analysis of stabilization diagrams and implemented it on a set of response measurements of a bridge. First, the authors used the SSI-COV method to build the stabilization diagrams of each signal, classifying as stable all poles whose modal parameters respect the limits of variation from one order to another. Considering the block row problem preciously discussed, the authors pointed out that a separate investigation was performed to deal with it. Then, hierarchical clustering was employed to group stable poles from the stabilization diagram. The clustering was performed with the single linkage algorithm and with a similarity measure that includes both the frequency difference and the MAC (Modal Assurance Criterion) value between a pair of modes. The threshold for this distance was manually determined through the analysis of the results. Finally, an outlier analysis based on a statistical technique was performed in order to remove the extreme values of damping within each cluster.

Reynders, Houbrechts, and De Roeck [6], on the other hand, proposed three automated steps to group the poles and applied them to response measurements collected from two different bridges. SSI-COV was used to create the stabilization diagram, and parameters such as the number of block rows and the maximum order of the stabilization diagram were manually selected. In the first step, the poles of the diagram were divided into two groups with the K-means clustering algorithm: certainly spurious modes and possibly physical modes. The K-means algorithm input was a feature vector containing as many relevant

single-mode validation criteria as possible, such as frequency, damping ratio, mode shape distance measures, modal phase collinearity, mean phase deviation, etc. All poles from the certainly spurious cluster were removed from the stabilization diagram and the remaining poles were evaluated using Hard Validation Criteria (HVC), i.e., all poles whose damping ratio is out of the permissible range and that do not have a complex conjugated pair are also removed from the stabilization diagram. In the second step, the hierarchical clustering was used to group the possible physical modes, using the average linkage algorithm. Similar to Magalhães, Cunha, and Caetano [5], a similarity measure based on the frequency difference and on the MAC value between a pair of poles was considered. The similarity measure threshold was determined with an automated procedure that considers the results obtained in the previous step. In the last step, the K-means algorithm was once again employed to separate clusters of physical poles from clusters of spurious poles, being the algorithm input the number of poles in each cluster, and the clusters with highest number of elements were chosen as the physical ones. Since outliers could still be present in the clusters, the authors choose to select the pole with median damping ratio value in the cluster to represent that cluster.

Neu et al. [7] pointed out that the algorithm proposed by Reynders, Houbrechts, and De Roeck [6] is limited to approximately real vibration modes and therefore limits the damping ratio range, a premise that is not always suitable to more complex valuated mode shape systems. To overcome this issue, the authors developed a new automatic algorithm that works without any user-provided thresholds and does not place any limitations on the damping ratio or the complexity of the system under analysis. As well as the other studies mentioned, the first step of this approach was to perform the SSI-DATA for numerous system orders. Then, mathematical poles were removed through a HVC based on the real and imaginary parts of each pole. The next step was to separate the poles in two groups: the probably physical poles and the certainly mathematical ones. In this attempt, the authors used a K-means clustering technique and proposed a consistent feature vector, applying transformation and normalization techniques to highly biased vectors. Then, the hierarchical clustering technique was employed in the probably physical group with the average linkage algorithm and using as similarity measure the relationship between the frequency difference and the MAC value between a pair of poles. The similarity measure threshold was determined from the probability distribution function of probable physical modes. In addition, repeated poles of the same order inside each cluster were located and all but one were removed based on their proximity to the cluster's centroid. The obtained clusters were separated into physical and mathematical clusters based on the number of poles in each one. Finally, the authors applied the modified Thompson Tau technique to remove outliers and the average natural frequency, damping ratio, and mode shape of each cluster were selected to represent it. The authors applied the proposed technique in measurements from a wind tunnel investigation with a composite cantilever, promoting the assessment of the algorithm's performance with a highly damped structure and low signal-to-noise ratio conditions.

Cardoso, Cury, and Barbosa [8] proposed an algorithm inspired by the ones presented by Magalhães, Cunha, and Caetano [5] and Reynders, Houbrechts, and De Roeck [6] and applied it to data from a numerical experiment, from a laboratory experiment of a simply supported beam and from dynamic tests of a bridge. According to the authors, the main contributions of the proposed methodology rely on an innovative similarity measure that leads to a symmetric dissimilarity matrix, additional modifications regarding filtering the spurious modes with damping and Mode Phase Collinearity (MPC) criteria, and a novel cluster regrouping technique.

More recently, automated identification of modal parameters that uses clustering techniques was studied by Fan, Li, and Hao [9], Wu et al. [10], and Mugnaini, Fragonara, and Civera [11], extending the application of the proposed algorithms on a steel frame structure, bridges, and a helicopter blade. The subject was also approached in studies

presented in the International Operational Modal Analysis Conference (IOMAC) of 2022 by Amer et al. [12], Priou et al. [13], and Dreher, Storti, and Machado [14].

All abovementioned papers employed machine learning techniques to the automated identification of modal parameters and demonstrated its relevance to current research. However, most presented research focused on the development and validation of automatic algorithms for civil structures, which exhibit different characteristics when compared to rotating machinery. To the authors' knowledge, no specific automatic algorithm for interpreting rotating machinery stabilization diagram has yet been studied.

Regarding the application of OMA in rotating machinery, this has been and still is a subject of great importance. Since rotating machines are exposed to periodic excitation, present nonlinear behavior, closely spaced modes, among other conditions that make the application of OMA a challenge, and several authors recently investigated the applicability of OMA in rotating machines. Brandt [15] developed two methods for harmonic removal, the Frequency Domain Editing (FDE) and the Order Domain Deletion (ODD) methods. Gres et al. [16–18] proposed and applied a method for harmonic removal based on orthogonal projection, applying it to experimental data from a plate and a ship in operation. Gioia et al. [19] and Peeters et al. [20], on the other hand, investigated a harmonic removal technique based on cepstrum analysis, applying it to the drivetrain of a wind turbine. More recently in IOMAC of 2022, Dreher, Storti, and Machado [21] proposed a method to identify both forward and backward modes of a rotor that appeared as closely spaced modes difficult to differentiate via traditional OMA by the use of directional coordinates. In the same conference, Zivanovic et al. [22] presented a novel approach to harmonic disturbance removal in single-channel wind turbine acceleration data by means of time-variant signal modeling.

These studies emphasize the importance given to the expansion of OMA's techniques to rotating machinery. Therefore, the objective of this work is to develop a new algorithm, based on the algorithms previously described that considers the different characteristics of rotating machinery, such as the presence of harmonics, outliers, the gyroscopic effect, and the complexity of the mode shapes, but still retains user friendliness. The main novelty of this work is the development of an algorithm that can identify the modal parameters related to the rotor, not the structure, which was not presented so far in the literature. The proposed algorithm is applied to two different datasets: response measurements of a test rig with a rotor supported by hydrodynamic bearings, and response measurements of a test rig with a rotor supported by rolling bearings, all under different operating and excitation conditions. The bearings were under healthy conditions for the generation of both datasets. Since rotating machines are also usually subjected to unideal excitation conditions with regard to OMA's premise of white noise excitation, this study evaluates whether the automatic OMA algorithm is adequate for the identification of the rotor's modal parameters under different excitation conditions, such as colored noise, tapping, lower sampling frequency, among others that will be further exposed, and which is another novelty of this work.

Section 1 presented the motivation for the development of this work, together with the literature review. An overview of the approach proposed for this work is presented in Section 2, along with a brief explanation of the SSI-DATA algorithm, the explanation of the algorithm proposed for automatic modal identification, and the description of both datasets used in this work. Section 3 presents the results obtained with the proposed approach and comparisons with methodologies previously proposed in the literature are pointed out. Finally, Section 4 presents the conclusions.

2. Materials and Methods

2.1. Overview of the Proposed Approach

The present work was organized according to the diagram presented in Figure 1, in which the dotted areas indicate a sequence of steps that was performed repeated times in order to generate the indicated results. First, the datasets are generated. Since the same

test rig is used to generate both datasets, the test setup is carried out in order to place the corresponding bearing (rolling or hydrodynamic) in the test rig. The operating condition is also defined, the rotor starts its operation, and the vibration signals corresponding to that setup and operating condition are collected. With the vibration signals of all setups and operating conditions, the acquisition of both datasets is completed. More information about the datasets is provided in Section 2.4.

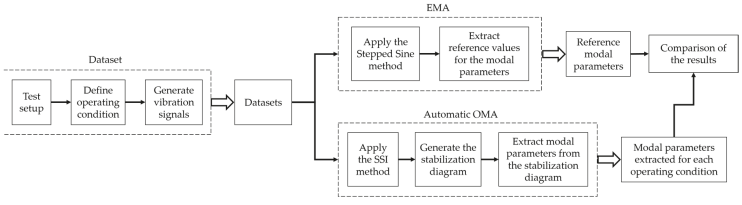


Figure 1. Diagram of the proposed approach.

With vibration signals of both datasets, EMA and OMA analyses are performed to identify the modal parameters of the system, that is the natural frequencies, damping ratios, and mode shapes. Both the EMA and OMA methods were applied to the vibration signals of the system in order to perform the identification. The EMA analysis is performed with the Stepped Sine method to determine reference values for the rotor's modal parameters. This is done for both the rotor supported by rolling and hydrodynamic bearings. An EMA analysis is also performed to determine the modal parameters of the rotor's foundation. More information about the EMA analysis is also provided in Section 2.4. The OMA analysis is performed as described in Section 2.3, using the automatic OMA algorithm proposed by this work. A brief summary of this section is the application of the SSI method in a set of vibration signals to generate a stabilization diagram. From each diagram, a series of stages (including machine learning techniques) extract the modal parameters from the system that originated the set of signals. The OMA analysis is performed for all setups and operating conditions. Finally, the modal parameters extracted from the automatic OMA are compared with the reference values, and the discussions are presented in Section 3.

2.2. Data Driven Stochastic Subspace Identification (SSI-DATA)

Although most of the papers presented in the previous section used the Covariance Driven SSI (SSI-COV) algorithm, there are indications that the Data Driven SSI (SSI-DATA) algorithm is more precise and robust [23,24]; thus, it was chosen in this research.

The Stochastic Subspace Identification is based on the stochastic model, defined by Equation (1):

$$\begin{cases} x_{k+1} = Ax_k + w_k \\ y_k = Cx_k + v_k \end{cases}, \tag{1}$$

in which $y_k \in \mathbb{R}^l$ denotes the outputs in the instant k, $x_k \in \mathbb{R}^n$ denotes the states in the instant k, and w_k and v_k denote the white gaussian noises, with zero mean, related to the process and the measurement noises, respectively. The white gaussian noises have the following covariance matrix:

$$E\left[\begin{pmatrix} w_p \\ v_p \end{pmatrix}\begin{pmatrix} w_q^T & v_q^T \end{pmatrix}\right] = \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}\delta_{pq}. \tag{2}$$

The system's order is n. Hence, the matrices dimensions are $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{l\times n}$, $Q \in \mathbb{R}^{n\times n}$, $S \in \mathbb{R}^{n\times l}$, and $R \in \mathbb{R}^{l\times l}$.

It is assumed that the pair $\{A, C\}$ is observable, which implies that all modes of the system can be observed in the outputs y_k and, therefore, can be identified. It is also assumed

that the pair $\left\{A, Q^{1/2}\right\}$ is controllable, which implies that all dynamic modes of the system are excited by the process noise.

The purpose of SSI is to use the outputs of the system to determine the systems matrices A and C, and, with them, extract the modal parameters of the system.

In order to do that, the first step is to build the output block Hankel matrix ($Y_{0|2i-1}$), that can be divided into the block Hankel matrices of the past outputs (Y_p) and the future outputs (Y_f) and is given by:

$$Y_{0|2i-1} \triangleq \left[\begin{array}{cccc} y_0 & y_1 & \cdots & y_{j-1} \\ \cdots & \cdots & \cdots & \cdots \\ y_{i-2} & y_{i-1} & \cdots & y_{i+j-3} \\ \hline y_{i-1} & y_i & \cdots & y_{i+j-2} \\ y_i & y_{i+1} & \cdots & y_{i+j-1} \\ y_{i+1} & y_{i+2} & \cdots & y_{i+j} \\ \cdots & \cdots & \cdots & \cdots \\ y_{2i-1} & y_{2i} & \cdots & y_{2i+j-2} \end{array}\right] = \left(\frac{Y_{0|i-1}}{Y_{i|2i-1}}\right) = \left(\frac{Y_p}{Y_f}\right), \tag{3}$$

in which i is the number of block rows and j is the number of block columns. Then, the projection matrix ($\mathcal{O}_i$) can be determined by the projection of the future outputs onto the past outputs and can be obtained through the QR decomposition of the output block Hankel matrix:

$$\mathcal{O}_i = Y_f / Y_p \tag{4}$$

The SVD decomposition is then applied to a product of the projection matrix and weighting matrices that are selected based on the desired algorithm (Principal Component Analysis—PCA, Unweighted Principal Components—UPC, or Canonical Variate Algorithm—CVA):

$$W_1 \mathcal{O}_i W_2 = USV^T = (U_1 U_2)\begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} = U_1 S_1 V_1^T. \tag{5}$$

The projection matrix can also be expressed as the product of the extended observability matrix (Γ_i) and the forward Kalman filter state sequence ($\hat{X}_i$):

$$\mathcal{O}_i = \Gamma_i \hat{X}_i. \tag{6}$$

Therefore, the extended observability matrix and the state sequence are determined by:

$$\Gamma_i = W_1^{-1} U_1 S_1, \tag{7}$$

$$\hat{X}_i = \Gamma_i^{\dagger} \mathcal{O}_i. \tag{8}$$

Similar operations can be used to determine the shifted state sequence ($\hat{X}_{i+1}$). Then, the system's matrices A and C can be determined applying the least square method to the following equation, derived from the stochastic model (Equation (1)):

$$\begin{bmatrix} \hat{X}_{i+1} \\ Y_{i|i} \end{bmatrix} = \begin{bmatrix} A \\ C \end{bmatrix} \hat{X}_i + \begin{pmatrix} \rho_w \\ \rho_v \end{pmatrix}, \tag{9}$$

in which ρ_w and ρ_v are the Kalman filter residues. The modal parameters extraction, along with more details about the hole procedure, can be found in [23].

2.3. Algorithm

The proposed automatic algorithm was divided in the following steps:

1. Create the stabilization diagram using the SSI algorithm and classify each pole based on stabilization criteria;

2. Clear the stabilization diagram using the Hard Validation Criteria (HVC);
3. Group poles that represent the same mode using agglomerative hierarchical clustering;
4. Remove from each cluster poles of repeated orders, so that only one pole of this order remains;
5. Eliminate small clusters that probably represent clusters of spurious or mathematical poles;
6. Perform an outlier detection based on the boxplot method;
7. Describe the global modes by the clusters mean frequency, mean damping, and mean mode shape;
8. Group poles with mode shapes of high correlation using agglomerative hierarchical clustering.

In the following, the choice of all above-mentioned steps is justified and clarified.

2.3.1. Stabilization Diagram and Stabilization Criteria

For the first step, it is possible to employ either SSI-COV or SSI-DATA algorithms, although the authors decided for the last. The identification is performed with increasing model orders and all extracted poles are inspected and classified according to the following evaluation. A k-order pole is stable if there is at least one (k−1)-order pole that satisfies the following stabilization criteria:

$$\Delta f_{m,n} = \frac{|f_n - f_m|}{f_n} < lim_f, \tag{10}$$

$$\Delta \zeta_{m,n} = \frac{|\zeta_n - \zeta_m|}{\zeta_n} < lim_\zeta, \tag{11}$$

$$MAC_{m,n} = \frac{\left|\varphi_m^H \varphi_n\right|^2}{(\varphi_m^H \varphi_m)(\varphi_n^H \varphi_n)} > lim_{MAC}, \tag{12}$$

where m corresponds to the pole of order k under evaluation and n corresponds to any pole of order $(k-1)$. All limits are manually selected, but suitable values can be easily determined through initial analyses of the system. All poles that do not fulfill the stabilization criteria are classified as not stable.

2.3.2. Hard Validation Criteria (HVC)

The idea behind the HVC is to remove all certainly spurious poles from the analysis of the following steps. In order to detect these poles, two criteria are employed: the damping ratio, and information about complex conjugated pairs.

As physical modes are characterized by positive damping ratios, it is expected that all poles with negative damping are spurious. Moreover, performing initial tests allow the analyst to know the normal behavior of the system, including the normal range of damping ratios. Thus, modes from the rotor or from its foundation are usually known within a determined range of damping. Furthermore, as already mentioned, rotating machines are constantly excited by periodic signals coming from their own operation or from the operation of other rotating parts in their surroundings. These harmonic frequencies appear in the stabilization diagram as stable poles of low or negative damping ratio due to their statistical aspects. Therefore, a first filter based on the poles damping ratio can be stablished as an HVC, all poles with values that are negative or out of the expected range being spurious, as employed by [6,8].

As mentioned by [6,7], every physical mode of a system appears in complex conjugated pairs, which makes it possible to classify as spurious all poles from the diagram that does not have a complex conjugated pair and remove them for the subsequent analysis.

2.3.3. Agglomerative Hierarchical Clustering

Before applying the hierarchical clustering, some of the papers mentioned in the introduction added a step that would separate certainly spurious poles from probably physical ones using some criteria based on the pole's mode shape complexity, such as Mode Phase Collinearity (MPC) and Mode Phase Deviation (MPD). As will be presented in the results, during the development of the algorithm proposed in this paper, it was observed that these criteria are not suitable to distinguish the rotor's modes in the stabilization diagram. In order to avoid criteria that are not suitable for rotating machinery, it was decided to not apply a step before the hierarchical clustering.

As with all papers presented in the previous section, the machine learning technique called agglomerative hierarchical clustering was selected to group poles that represent the same mode. According to [7], the average linkage showed better results to create compact clusters of individual physical modes, thus it is used as the algorithm for hierarchical clustering. Moreover, in this paper, it was decided to employ the same algorithm with a similarity measure based solely on the frequency difference between all pair of poles, and the analysis of the MAC value within each cluster is postponed at the end of the algorithm.

$$\Delta f_{m,n} = \left| \frac{f_n - f_m}{\max(f_n, f_m)} \right|. \tag{13}$$

Since only the frequency difference is used as similarity difference, the threshold can be easily selected. This can be done from an analysis of the modes of interest variation in the stabilization diagram.

2.3.4. Removal of Poles from Repeated Orders

In the case of closely spaced modes or spurious and mathematical poles near physical poles, it can be that more than one pole from the same order are grouped in the same cluster, which is not appropriate since each cluster is supposed to represent a single physical mode. Aiming to remove the repeated poles, a comparison of the damping ratio of each repeated pole with the cluster's damping ratio median is done, given that no outlier removal was yet performed, and only the repeated pole with damping ratio closest from the median is maintained.

2.3.5. Small Clusters Removal

Since physical modes tend to have a better stabilization when compared to spurious and mathematical poles (i.e., appear at several model orders in the stabilization diagram), the number of poles in each cluster can be used to separate clusters of spurious or mathematical poles from clusters of physical poles. A study by [7] presents a methodology that eliminates all clusters with sizes lower than 50% of the biggest cluster size. However, stabilization diagrams of rotating machinery data comprise both structural and rotor modes, the last being usually harder to stabilize in comparison with the first. Therefore, a 50% limit proved to exclude some rotors' modes of interest from the analysis and the mean size of all clusters was adopted as a threshold.

2.3.6. Outlier Detection

As a result of adopting just the frequency as a measure of similarity, a possible effect is that poles with different damping factors are grouped together in one cluster. As will be presented in the results, the SSI method is usually able to identify one of the closely spaced modes of the rotor (backward or forward) with an acceptable range of damping, whereas the other mode is identified with a lower or higher (or simply different) damping ratio. Given that these modes are closely spaced, it is possible that they end up grouped in the same cluster. In order to eliminate the modes with lower or higher (or simply different) damping, the outlier detection proposed by [5] is adopted. This approach was chosen because of the lack of information about the probability distribution of the clusters, and

the outlier detection based on the quartile's information results in a more conservative and effective method to remove outliers.

Furthermore, it is possible that a mode is identified in the stabilization diagram with high dispersion or with poles that, due to the order, end up far from the average of the mode. Aiming to maintain clusters with low frequency dispersion, the same approach adopted to detect damping ratio outliers is considered to detect frequency outliers. Thus, the outlier analysis is performed for both damping and frequency values.

2.3.7. Global Modes

Finally, each cluster mean frequency, mean damping ratio, and mean mode shape are extracted to describe the global modes. The mean was adopted because an outlier analysis was performed, but it is also possible to use the pole with median damping ratio, as done by [6].

2.3.8. Agglomerative Hierarchical Clustering of Each Cluster

Since the MAC value was not employed in the similarity measure of the third step and the mode shapes of each cluster were not evaluated in any other step of the algorithm, it is possible that poles with inaccurate mode shapes were included in the results, which would render the mean mode shape estimate also inaccurate. In order to remove these poles from the clusters, the hierarchical clustering algorithm can be once again employed as an additional step of the algorithm to improve the estimates, as described above.

In order to implement this step, the MAC value is computed between all poles within each cluster, resulting in one MAC matrix for each global mode extracted by the algorithm. The minimum value of each matrix is then identified and compared with the MAC limit of the stabilization diagram, informed by the user in the first step of this algorithm. If the minimum value of a cluster's MAC matrix is above the limit ($MAC_{min} > lim_{MAC}$), it means that all mode shapes within this cluster have high correlation and, therefore, that the mean mode shape computed in the last step is adequate to represent the mode shape of that global mode. However, if the minimum value of a cluster's MAC matrix is below the limit ($MAC_{min} < lim_{MAC}$), it means that not all mode shapes of this cluster have high correlation and that the mean mode shape is not adequate to represent the cluster. In this last case, another processing step is required to remove the poles with low correlation and obtain another set of poles with mode shapes that have high correlation between them and that can represent the mode shape of that global mode. In order to do that, hierarchical clustering is employed with a similarity measure equal to the inverse of the MAC between two poles of the global mode under analysis:

$$\Delta MAC_{m,n} = \frac{1}{MAC_{m,n}}. \tag{14}$$

This way, the two poles that have low correlation (low MAC value) will be distant from each other, whereas the two poles that have high correlation (high MAC value) will be close to each other. For the threshold value, the inverse of the MAC limit of the stabilization diagram is employed, so that the resulting clusters will comprise only the poles with MAC values above the limit. Then, the biggest cluster is identified and only the poles from this cluster are selected to represent the global mode.

Once this procedure is performed for all clusters from the previous step, the means of the frequencies, damping ratios, and mode shapes are once more computed to represent each global mode.

The resulting algorithm is summarized in Algorithm 1.

Algorithm 1: Proposed Algorithm.

Inputs: Stabilization diagram (frequency, damping ratio, and mode shape), damping ratio limits (ζ_{min} and ζ_{max}), stabilization criteria (lim_f, lim_ζ and lim_{MAC}), and similarity measure threshold (lim_D).

Algorithm 1: Proposed Algorithm. *Cont.*

Output: Global modes

1. Classify as stable all poles that satisfy the stabilization criteria and as not stable all remaining poles
2. Classify as spurious all poles with damping ratio lower than ζ_{min} or higher than ζ_{max} (Hard Validation Criteria—HVC) or that do not appear with a complex conjugated pair
3. Extract the number of stable poles (n_{me})
4. Create a matrix of zeros $D \in \mathbb{R}^{n_{me} \times n_{me}}$
5. For m in $[1, n_{me}]$:
   - 5.1. For n in $[1, n_{me}]$: Compute the distance between the poles m and n ($d_{m,n}$) using the relative distance between the natural frequencies of both poles and assign the result to the matrix D in the position (m, n)
6. Apply agglomerative hierarchical clustering taking the distance matrix D as the method's similarity measure and consider the informed threshold (lim_D)
7. Extract the number of clusters obtained (n_c)
8. For c in $[1, n_c]$:
   - 8.1. If cluster c has more than one pole of each order, remove all poles of each order but one, and keep the one with the damping ratio closest to the cluster's damping ratio median
   - 8.2. Store the number of poles and each modal parameter (natural frequency, damping ratio, mode shapes and order) of the cluster c
9. Create a histogram of the number of poles in each cluster
10. Extract the mean size of the clusters
11. Select the clusters whose size is bigger than the mean size
12. Create a boxplot of the frequency and of the damping ratio
13. Remove the outliers:
    - If $\omega_n < Q_{1\,freq} - 1.5\ IQR_{freq}$ or $\omega_n > Q_{3\,freq} + 1.5\ IQR_{freq}$, remove the pole n because it is a frequency outlier
    - If $\zeta_n < Q_{1\zeta} - 1.5\ IQR_{\zeta}$ or $\omega_\zeta > Q_{3\zeta} + 1.5\ IQR_{\zeta}$, remove the pole n because it is a damping ratio outlier

    Being Q_1 is the first quartile, Q_3 the third quartile, and IQR the difference between the upper and lower quartiles
14. Extract the parameters that represent the clusters: mean frequency, mean damping ratio, and mean mode shape
15. Extract the number of global modes (n_{gm})
16. For i in $[1, n_{gm}]$:
    - 16.1. Extract the number of poles (n_p)
    - 16.2. Create a matrix of zeros $D_{MAC-i} \in \mathbb{R}^{n_p \times n_p}$
    - 16.3. For m in $[1, n_p]$:
      - 16.3.1. For n in $[1, n_p]$:

      Compute the MAC value between the poles m and n and assign the result to the matrix D_{MAC-i} in the position (m, n)
    - 16.4. Extract the minimum value of the matrix D_{MAC-i} (min_i)
    - 16.5. If $min_i < lim_{MAC}$:
    - 16.6. Create a matrix of zeros $D_i \in \mathbb{R}^{n_p \times n_p}$
    - 16.7. For m in $[1, n_p]$:
    - 16.8. For n in $[1, n_p]$:
    - 16.9. Compute the distance between the poles m and n according to Equation (14) and assign the result to the matrix D_i in the position (m, n)
    - 16.10. Apply agglomerative hierarchical clustering taking the distance matrix D_i as the method's similarity measure and considering the informed MAC limit ($1/lim_{MAC}$)
    - 16.11. Select the poles from the biggest cluster to represent the global mode i
    - 16.12. Extract the parameters that represent the modal globe: mean frequency, mean damping ratio, and mean mode shape

2.4. Description of Datasets

2.4.1. Test Rig with Hydrodynamic Bearings

The first data set used in this work was taken from a test rig with a rotor supported by hydrodynamic bearings, displayed on Figure 2. The system is basically composed of a rotating steel shaft (15 mm in diameter and 719 mm in length) supported by two hydrodynamic bearings (31 mm diameter, 18 mm length, 90 µm of radial clearance, and ISO VG32 oil at ambient temperature as working fluid) connected to an electric motor through a flexible coupling. In addition, the system has a hard disk and an electromagnetic actuator (used to insert different types of noise into the rotor). The experiments were carried out with the rotor operating with an angular shaft velocity of 75 Hz and four accelerometers installed in both bearings (two accelerometers for each bearing) were used to collect the vibration on the Y and Z directions.

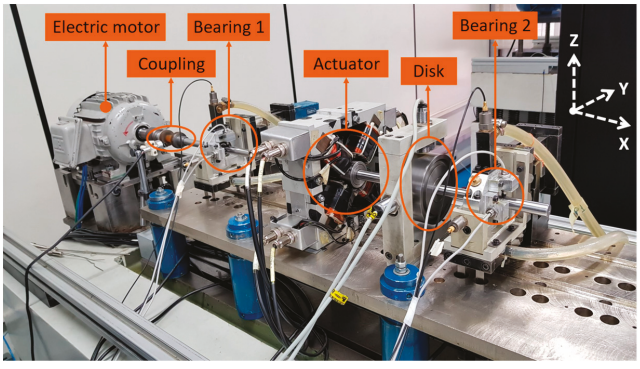


Figure 2. Test rig with hydrodynamic bearings.

During operation, rotating machines can be subjected to different types of excitation conditions that can facilitate or hinder OMA's application. In order to investigate it, Ref. [25] performed the identification of a rotating system through OMA techniques and revealed that different test conditions influence the extracted parameters, ranging from non-identification to precise identification of modal parameters, which characterizes challenges to the automatic algorithm proposed here. Hence, it was decided to use more than one test condition. For that, the data was collected with different inputs and sampling frequencies, and during different periods of time, resulting in the tests displayed in Table 1.

Table 1. Test conditions for the test rig with hydrodynamic bearings.

Test	f_s [Hz]	Time [s]	Excitation Direction	Excitation
1	2048	240	Y	White noise—medium intensity
2	2048	240	Y	White noise—low intensity
3	2048	240	Z	White noise and tapping
4	2048	240	Y	Blue noise
5	1024	240	Y	White noise—medium intensity
6	2048	120	Y	White noise—medium intensity

An EMA analysis was also carried out though the Stepped Sine method to determine the modal parameters of the rotor supported by hydrodynamic bearings, so that their correct values were known for further validation of the proposed OMA algorithm. For this test, the rotor's speed was 75 Hz. Two sets of tests were carried out with a step of 0.25 Hz, the first one with frequency range between 48 Hz and 58 Hz, in order to identify the first rotor's mode, and the second one with frequency range between 200 Hz and 220 Hz, in order to identify the second rotor's mode. To each test, 5 measurements were collected to

compute mean values and diminish random errors. The results are displayed in Table 2. It is important to emphasize that the Stepped Sine method was able to identify two pairs of natural frequencies, each one containing the forward and the backward frequencies of the rotor, whose occurrence is traced back to the gyroscopic effect.

Table 2. Modal parameters of the rotor supported by hydrodynamic bearings.

Mode	Backward		Forward	
	Freq. [Hz]	**Damp. [%]**	**Freq. [Hz]**	**Damp. [%]**
First	52.8	4.26	53.1	4.25
Second	212.6	2.45	212.2	2.48

During the experiments, it was found that modal information of the foundation was transferred to the rotor's dynamic response. A further modal analysis of the foundation was required so that the modal parameters extracted through OMA could be properly assigned to the system component that originated them. For the extraction of the foundation's modal parameters, EMA was applied to the foundation after the shaft removal and with the use of FRF estimators and an impact hammer. The structure's excitation was performed by means of impulses applied to the bearing housings in the Y and Z directions and the responses were measured using accelerometers mounted in the three directions (X, Y and Z) of the bearing housings. Frequency Response Functions (FRFs) were estimated, gathered, and evaluated only in the frequency range of interest (80 Hz to 320 Hz). The Least Square Complex Exponential (LSCE) algorithm was employed to estimate the modal parameters and the results are depicted in Table 3.

Table 3. Foundation's modal parameters.

Mode	Freq. [Hz]	Damp. [%]
1	101.8	4.9
2	110.4	6.4
3	114.5	3.4
4	124.8	2.7
5	133.7	3.9
6	138.6	4.3
7	157.4	6.1
8	179.7	1.7
9	196.0	3.3
10	204.0	1.7
11	241.9	2.7
12	277.2	1.6
13	299.8	0.9

It is important to mention that although several foundation modes were identified, not all of them are excited during the rotor's operation, which causes only a few to appear when applying modal analysis through the rotor's vibration signals.

2.4.2. Test Rig with Rolling Bearings

The second data set employed in this work was taken from the same test rig presented in Figure 2, replacing the hydrodynamic bearings by rolling bearings (15 mm inner diameter NJ 202 by NSK®) and using different excitation conditions. There are only minor variations in the positioning of each component due to the inherent inaccuracy of the assembly, disassembly, and alignment process of the system. The goal of these tests was also to evaluate the proposed algorithm in a system with lower damping, as expected for rolling bearings when compared to hydrodynamic bearings. Four accelerometers installed in both bearings (two accelerometers for each bearing) were again employed to collect the vibration

on the Y and Z directions. The experiments were carried out with the rotor rotating in 30 Hz and under different operating conditions, resulting in the tests displayed in Table 4.

Table 4. Test conditions for the test rig with rolling bearings.

Test	f_s [Hz]	Time [s]	Excitation Direction	Excitation
1	1024	60	Y	White noise—medium intensity
2	1024	60	Y	White noise—low intensity
3	1024	60	Y	White noise and tapping
4	1024	60	Y	Blue noise

In order to evaluate OMA's results, an EMA analysis was also carried out through the Stepped Sine method. For this test, the rotor's speed was 30 Hz, and the test was carried out with frequency range between 20 Hz and 75 Hz and a step of 0.25 Hz, with the aim of evaluating only the first vibrating mode of the system. Two tests were carried out, one where the excitation was applied in the Y direction and other where the excitation was applied in the Z direction. The results are displayed in Table 5, where the values correspond to the obtained averages.

Table 5. Modal parameters of the rotor supported by rolling bearings.

Mode	Backward		Forward	
	Freq. [Hz]	Damp. [%]	Freq. [Hz]	Damp. [%]
First	51.35	1.168	52.65	0.864

As before, the Stepped Sine method was able to identify a pair of natural frequencies, containing the forward and the backward frequencies of the rotor. The significant reduction in damping values is noted when compared to the system supported by hydrodynamic bearings (compare Tables 2 and 5). Regarding the small variations in the natural frequencies, these are more related to the inherent difficulty of positioning the components, as previously mentioned.

3. Results

The proposed algorithm is applied to two different datasets: response measurements of a test rig with a rotor supported by hydrodynamic bearings, and response measurements of a test rig with a rotor supported by rolling bearings.

The results obtained through the test rig with the rotor supported by hydrodynamic bearings are the first ones to be presented. To illustrate all steps of the algorithm, clarifying the analyzes performed by them, test 1 of Table 1 is taken as the standard example and a comprehensive explanation of its results is presented. Then, the algorithm is applied to all other tests in Table 1 and the main results are presented and discussed in order to show the algorithm's robustness when different operating conditions are present.

Later, the test rig supported by rolling bearings, which has a higher stiffness and a lower damping when compared to the first test rig, is analyzed to verify the algorithm's robustness to distinct systems. The results of all tests of Table 4 are briefly presented and discussed.

The algorithm, as well as the SSI-DATA method, were implemented in the programming language Python™.

3.1. Test Rig with Hydrodynamic Bearings

The stabilization limits considered in the following analysis were 0.2% for the frequency variation, 2% for the damping ratio variation, and 95% for the minimum MAC value, all of them conservatively chosen. The range $[0.3\%, 10\%]$ was used as the damping ratio limit. All stabilization diagrams were built with a maximum order of 100, with fixed 100 block rows.

Figure 3 displays the stabilization diagram of the first test of Table 1, excitation with white noise (medium intensity) and a sampling frequency of 2048 Hz. The diagram is presented in the frequency range of 0 Hz to 256 Hz, the range of interest in this analysis. From the diagram, one can observe that there are three alignments of spurious poles, the first at 75 Hz (the rotor's rotating speed), two at 150 Hz (first harmonic), and the last at 225 Hz (second harmonic). The identification of the rotating speed and its harmonics as spurious was possible due to the HVC related to the damping ratio. In addition, several mathematical poles were also classified as spurious and, therefore, will not enter the following analysis. One can also observe that, close to the first rotor's mode, two poles are predominantly identified in each order, which could lead to the idea that both forward and backward frequencies are identified. However, the second poles of each order are mostly identified with a high damping ratio (>7%), being inadequate to represent any rotor's frequency.

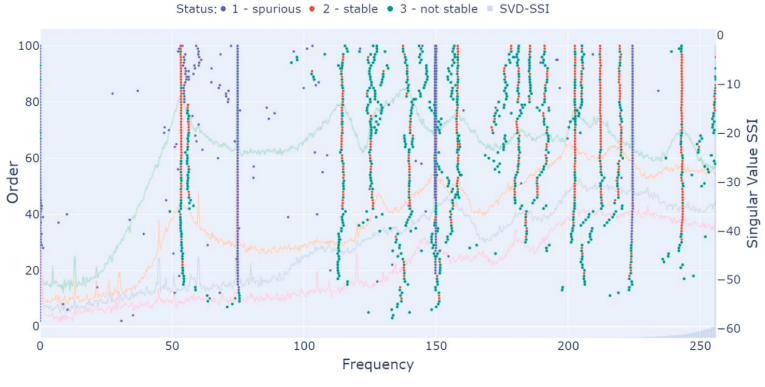


Figure 3. Test 1 (white noise—medium intensity) stabilization diagram.

The MPC (computed as described in [26]) and the MPD (computed as described in [6]) values of each pole were computed to perform additional analysis. The MPC value ranged from 63% to 99% for the first rotor's mode, the highest ones (>86%) being outliers because of the high damping ratio (>5%), as will be seen in a further outlier analysis. For the second one, the range was 98% to 100%. For the foundation mode of 241.9 Hz, the values were much more stable, ranging from 94% to 98%. The MPD value, in contrast, ranged from 8% to 35% for the first rotor's mode, the lowest ones (<19%) being outliers because of the high damping ratio (>5%). For the second one, the range was 3% to 6%. For the foundation mode of 241.9 Hz, the range was 11% to 16%. Therefore, if any clustering algorithm or HVC based on the MPC or MPD values were employed, the first rotor's mode could be identified due to its great dispersion as spurious, and the identification algorithm would fail to provide reliable information.

After building the stabilization diagram and applying the HVC, the hierarchical clustering was performed. For the selected threshold definition, the distance between the known difference of closely spaced modes was employed. The difference between the first and second frequencies of the first mode, according to Equation (13), is 0.006. For the second mode, the difference is 0.002. Tests considering thresholds near these values were evaluated, resulting in a selected threshold of 0.01. It is important to emphasize that this threshold proved itself adequate for all other tests of Table 1, demonstrating how simple it is to select a value that works in different operating conditions of the same system. Figure 4 displays the obtained dendrogram, in which each cluster is represented by a different color in the bottom of the dendrogram and whose x-axis is organized with the frequency range of 53 Hz to 250 Hz, distributed in an ascending order.

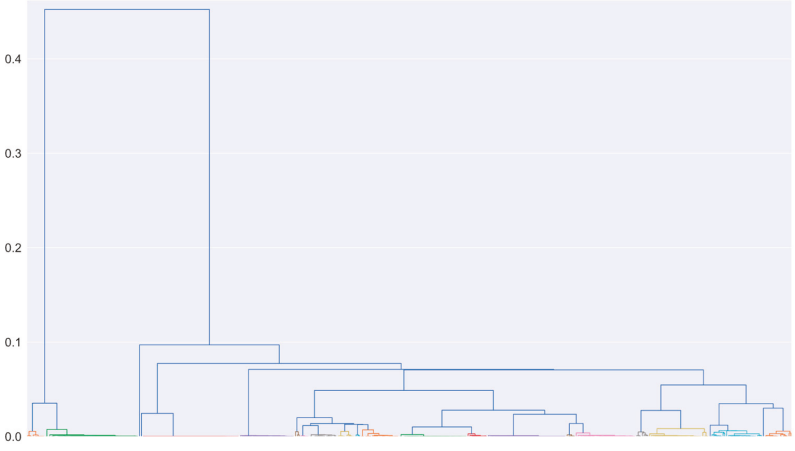


Figure 4. Test 1 (white noise—medium intensity) hierarchical clustering dendrogram.

Figure 5 displays the diagram of each cluster's size, along with the limits proposed by this paper and by [7] to remove small clusters. From Figure 5, one can see that if the limit proposed by [7] was considered, the 6th and the 8th foundation modes would not be identified by the algorithm. There are also cases in which the first rotor's mode is below the limit proposed by the authors, as the signals obtained from test 3 show. Therefore, the limit defined by the mean size is justified.

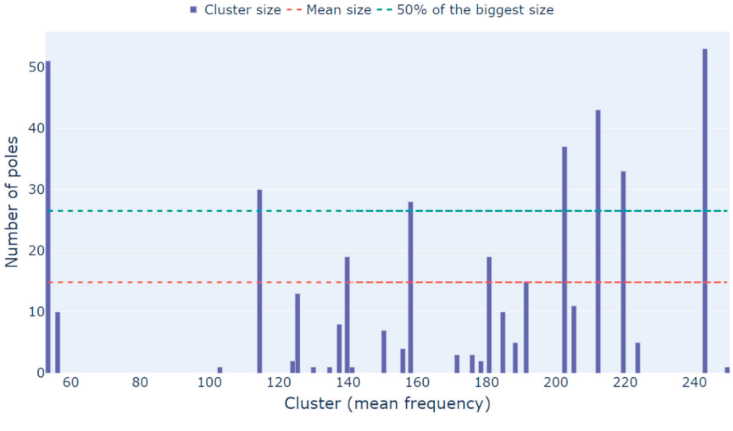


Figure 5. Test 1 (white noise—medium intensity) small clusters removal.

The outlier analysis was performed within the 10 clusters that remained from the previous analysis. Figure 6 displays the boxplot of both frequency and damping ratio values. Points out of the box range are considered outliers. Taking the first cluster as an example, which represents the first rotor mode, there are outliers in both frequency and damping ratio, although the first ones (53.16 Hz, 53.24 Hz, and 53.94) are less pronounced than the last ones (all damping ratios above 4%). From Figure 6, it is possible to see that the outlier analysis was adequate to remove outliers from all modes.

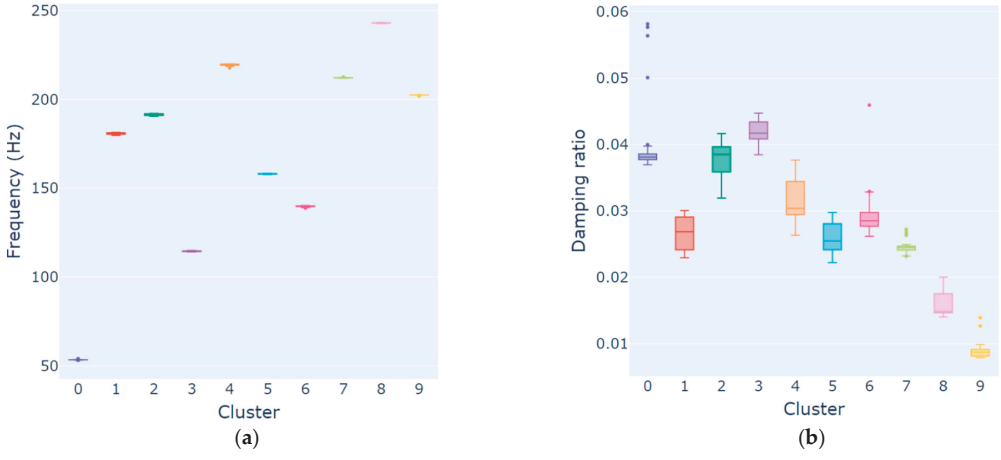


Figure 6. Test 1 (white noise—medium intensity) outlier analysis for frequency (**a**) and damping (**b**).

Concluding all essential steps proposed by the algorithm, the averages of the frequencies and of the damping ratios of the poles inside each cluster are extracted. The results are displayed in Table 6, along with the standard deviation of these parameters, the difference between the maximum and minimum values within the cluster that originated them, the errors in relation to the EMA references, the size of the cluster, and the lowest value in the MAC matrix, which will be further employed in the optional step to obtain sets of poles with high correlation mode shapes. From Table 6, one can see that most of the identified modes presented low standard deviations and low differences between maximum and minimum, for both frequency and damping ratio, and bigger cluster sizes.

Table 6. Test 1 (white noise—medium intensity) test rig with hydrodynamic bearings global modes.

f [Hz]				ζ [%]				Size	MAC (Minimum)
Mean	**Std.**	$\Delta_{max,min}$	**Error**	**Mean**	**Std.**	$\Delta_{max,min}$	**Error**		
53.4	0.05	0.2	0.56%	3.81%	0.06%	0.28	11.55%	43	98%
114.5	0.26	0.9	0.00%	4.19%	0.18%	0.63	18.85%	30	99%
139.9	0.21	0.7	0.93%	2.84%	0.11%	0.37	51.41%	16	99%
158.1	0.18	0.7	0.44%	2.60%	0.23%	0.75	134.62%	28	73%
180.8	0.62	1.8	0.61%	2.65%	0.25%	0.71	35.85%	19	83%
191.5	0.53	1.5	2.35%	3.78%	0.30%	0.97	12.70%	15	91%
202.6	0.07	0.2	0.69%	0.87%	0.05%	0.20	95.40%	35	98%
212.2	0.10	0.4	0.00%	2.44%	0.04%	0.15	1.64%	37	99%
219.5	0.40	1.5	-	3.11%	0.26%	0.92	-	32	90%
243.1	0.16	0.7	0.49%	1.58%	0.18%	0.60	70.89%	53	97%

It is important to mention that, although the first two modes of the rotor are composed by two frequencies, the backward and the forward ones (Table 2), the algorithm was not able to identify both of them. Since the similarity measure encompasses only the frequency difference between the poles, as presented in Equation (13), and considering the fact that the frequency and the damping ratio of the backward and forward frequencies are significantly close, it would be possible that both frequencies were grouped in the same cluster. However, the minimum MAC value for this mode was 98%, indicating a high correlation between all mode shapes within the cluster. Since some difference is expected from the mode shapes of the forward and backward frequencies, it is more likely that only poles of one of these frequencies are present in the cluster, indicating that the proximity of these two frequencies lead the SSI method to identify only one of them.

It is also important to mention that not only the rotor's modes were identified, but also several modes from the foundation. Comparing Table 6 with Table 3, one can see that the modes identified with the OMA algorithm do not have the exact same parameters as the modes identified by EMA (but are relatively close). However, one must also recall that the EMA test was performed without the shaft and this variation of the modal parameters was already expected. Comparing the foundation's results with the rotor's results, one can observe that the errors were similar, highlighting the algorithms' ability to extract accurate modal parameters for both the rotor and the foundation.

Moreover, Table 7 displays the errors between the EMA values and estimated values of the rotor's modal parameters using the proposed algorithm, in which all but one parameter presented a low error. The highest error was on the damping factor of the first mode, whose occurrence can be traced back to the SSI method's ability to estimate this parameter. Tables 6 and 7 demonstrate the proposed algorithm's capability of extracting the modal parameters of a rotating machine.

Table 7. Test 1 (white noise—medium intensity) rotor mode's error.

Parameter	First Mode			Second Mode		
	EMA	**OMA**	**Error**	**EMA**	**OMA**	**Error**
f [Hz]	52.8	53.4	1.14%	212.6	212.2	0.19%
	53.1		0.56%	212.2		0.00%
ζ [%]	4.26	3.81	10.56%	2.45	2.44	0.41%
	4.25		10.35%	2.48		1.61%

With the clusters of each global mode and the lowest value in their MAC matrices, the additional step of the algorithm can be performed. The modes of 53.4 Hz, 114.5 Hz, 139.9 Hz, 202.06, 212.2 Hz, and 243.1 Hz presented good results, since the minimum values on their MAC matrix were greater than the MAC limit of the stabilization diagram (95%), an expected value from poles from the same mode. Therefore, no alteration will be performed in the clusters of these modes. However, the other modes (158.1 Hz, 180.8 Hz, 191.5 Hz and 219.5) presented values lower than the MAC limit of the stabilization diagram. Hence, hierarchical clustering based on the MAC values was performed, obtaining, for each mode, a new set of poles from which the mean, the standard deviation, and the difference between the maximum and minimum values of the modal parameters were computed. The results are displayed in Table 8, from which one can verify that the minimum MAC value of all modes is now at least 95%, indicating that the obtained clusters present mode shapes with high correlation and, therefore, the mean of the mode shapes of each cluster is adequate to represent these modes. It is also possible to verify that no significant alteration occurred on the mean values of the modal parameters. In addition, the standard deviation and the difference between the maximum and minimum of most of the clusters achieved lower values (values highlighted in green), whereas only two modes exhibited an increase in standard deviation (values highlighted in red).

Table 8. Test 1 (white noise—medium intensity) test rig with hydrodynamic bearings global modes after hierarchical clustering based on the MAC value.

f [Hz]			ζ [%]			**MAC (Minimum)**
Mean	**Std.**	$\Delta_{max,min}$	**Mean**	**Std.**	$\Delta_{max,min}$	
157.9	0.17	0.5	2.61%	0.13%	0.38	96%
180.6	0.81	1.8	2.62%	0.12%	0.41	95%
190.9	0.34	0.8	3.73%	0.18%	0.38	97%
219.6	0.52	1.5	3.28%	0.23%	0.64	95%

After these analyzes, the proposed algorithm, ignoring the additional step, was applied to all tests of Table 1 and the results obtained for the rotor's modes are displayed in Table 9.

As Table 9 shows, the proposed algorithm was able to extract the rotor's modes from all tests, these having a small standard deviation and with mean values close to the values selected via EMA. It is important to mention that the main reason for the high errors in the damping ratio estimations is the low magnitude of this parameter. Moreover, the estimation of damping ratios is a challenge even when well consolidated EMA techniques are used for the modal identification, and high errors are also obtained when the results of different EMA techniques are compared. In this context, the estimations displayed in Table 9 are very good.

Table 9. Rotor's global modes for the test rig with hydrodynamic bearings.

Mode	First Mode						Second Mode					
	f [Hz]			ζ [%]			f [Hz]			ζ [%]		
Test	**Mean**	**Std.**	**Error**	**Mean**	**Std.**	**Error**	**Mean**	**Std.**	**Error**	**Mean**	**Std.**	**Error**
EMA (backward)	52.8	-	-	4.26	-	-	212.6	-	-	2.45	-	-
EMA (forward)	53.1	-	-	4.25	-	-	212.2	-	-	2.48	-	-
1	53.4	0.05	0.56%	3.81	0.06	10.35%	212.2	0.10	0.00%	2.44	0.04	1.61%
2	53.7	0.09	1.13%	3.94	0.10	7.29%	211.9	0.05	0.14%	2.34	0.03	5.65%
3	52.5	0.09	1.13%	3.63	0.12	14.59%	211.6	0.15	0.28%	2.64	0.05	6.45%
4	53.1	0.05	0.00%	3.52	0.08	17.18%	211.8	0.19	0.19%	2.37	0.07	4.44%
5	52.9	0.12	0.38%	3.77	0.08	11.29%	211.8	0.11	0.19%	2.44	0.04	1.61%
6	53.1	0.02	0.00%	3.48	0.04	18.12%	211.9	0.15	0.14%	2.48	0.06	0.00%

As occurred in Test 1, the application of the proposed algorithm to the remaining tests of Table 1 also enabled the identification of several foundation modes. In order to summarize the results, Figure 7 displays all modes estimated through the proposed algorithm as black dots, all rotor modes as continuous lines, and all foundation modes estimated by EMA as dashed lines. The frequency is presented in the x-axis and the data used to estimate the modes is presented in the y-axis. As indicated by Figure 7, most foundation modes were identified. Recalling the stabilization diagram of Figure 3, obtained with the data with medium intensity white noise, one can see that there are some frequency ranges in which the stabilization is irregular. Therefore, the absence of some foundation modes can be, once more, associated with the challenges in the SSI method.

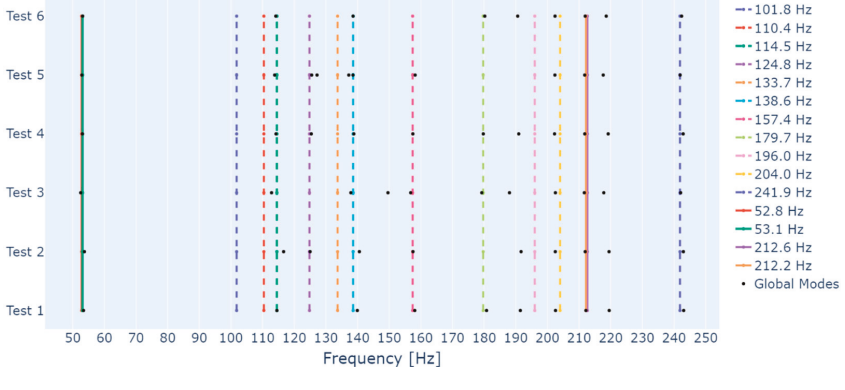


Figure 7. Foundation's modes identified through the proposed algorithm.

With these analyses, an investigation was performed to evaluate the differences between dividing the hierarchical clustering in two steps, one based only on the frequency difference between the poles, and other based only on the MAC value, as proposed in this paper, and applying the hierarchical clustering in one single step, considering both the frequency difference and the MAC value, as proposed by other papers in the literature.

In this case, the third step of the algorithm was modified so that the similarity measure comprised the frequency difference and the MAC value. Then, it was applied to all tests of Table 1, without the additional step, and considering four different threshold values (0.04, 0.06, 0.08 and 0.1). The results are displayed on Figure 8, along with the results from the proposed algorithm with the additional step to facilitate the comparison. In some cases, the modified algorithm identified global modes with very close frequencies. Due to the frequency range of Figure 8, these cases would not be visible. Therefore, the icons representing them have been modified, and are represented with solid icons rather than hollow ones.

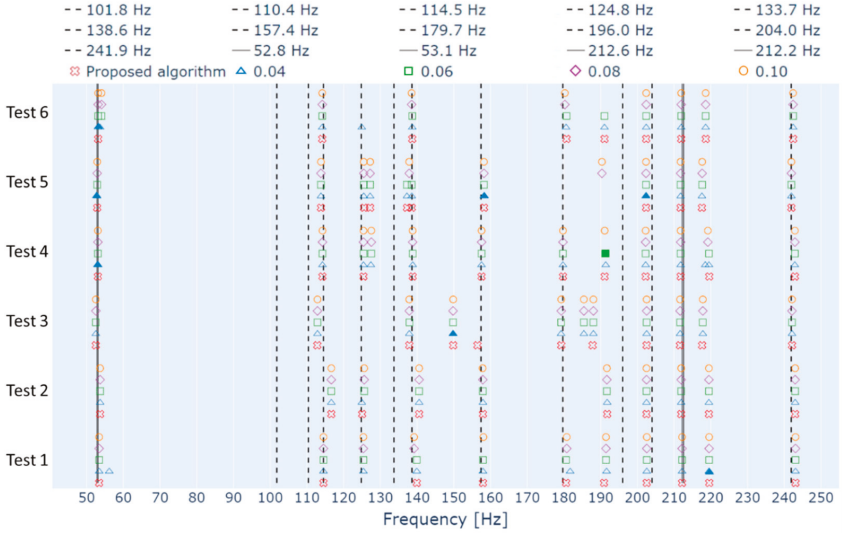


Figure 8. Foundation's modes identified through the proposed and modified algorithms.

From Figure 8, one can see that most of the frequencies identified by the proposed algorithm were also identified by the modified one. However, there are several cases in which two very close frequencies are identified, especially when the threshold of 0.04 is used. Analyzing the frequency range of the first rotor;s mode, one can see that the threshold of 0.04 identified two frequencies of approximately 53 Hz for Tests 1, 4, 5, and 6, and the thresholds of 0.06, 0.08, and 0.1 performed the same for Test 6. When the stabilization diagram of Figure 3 was analyzed, it was verified that this frequency range indeed presented the stabilization of two different modes. However, the damping factor of one of them made it inadequate to represent any rotor's frequency. That is also the case for all other tests. Therefore, the identification of two frequencies near the rotor modes by the modified algorithm represents a disadvantage of using one single hierarchical clustering with similarity distance that comprises both the frequency difference and the MAC value.

Evaluating other frequency ranges, it is possible to identify the same phenomenon in some foundation modes (124.8 Hz, 138.6 Hz, 157.4 Hz, 196.0 Hz and 204.0 Hz), mostly in the results from the modified algorithm (only the foundation modes of 124.8 Hz and 138.6 Hz of test 5 for the proposed algorithm). Analyzing each stabilization diagram, it was observed that most pairs of close frequencies were identified because poles from a single physical mode happened to be divided into more than one cluster by the algorithms due to irregularities in the stabilization diagram. The exceptions were the frequencies near 124.8 Hz of Tests 4 and 5, since the stabilization diagrams of these tests really present the alignment of two modes. However, it is possible that one of the alignments is actually an alignment of spurious modes rather than a closely spaced mode of the foundation, as occurred for the first rotor's frequency.

Furthermore, there are some cases in which a foundation mode was identified by one of the algorithms and not by the other. These cases occurred 16 times, for both algorithms and all thresholds, and occurred for the foundation modes of 124.8 Hz (Tests 1 and 6), 157.4 Hz (Tests 1 and 3), and 196.0 Hz (Tests 4, 5 and 6). In five of these cases, the employed algorithm was the modified one with a threshold of 0.08. The modified algorithm with thresholds of 0.06 and 0.10 were responsible for three cases each, and the modified algorithm with a threshold of 0.04 was responsible for two cases. The proposed algorithm, in turn, was responsible for three cases.

Moreover, when the modified algorithm is employed, there is no guarantee that the minimum MAC value between the poles of a global mode is above the limit established for the stabilization diagram. Considering the global modes identified in all tests, when the threshold of 0.04 is used, 7 of the 74 identified global modes presented MAC values below 95%, with the minimum being 91%. When the threshold of 0.06 is used, 21 of the 66 identified global modes present values below 95%, with a minimum of 88%. When the threshold of 0.08 is used, 30 of the 62 identified global modes present values below 95%, with a minimum of 80%. Finally, when the threshold of 0.10 is used, 32 of the 64 identified global modes present values below 95%, with a minimum of 80%.

Considering the results presented here and that only one threshold value was selected for all tests of the proposed algorithm, some findings must be summarized. When the modified algorithm with low thresholds is used, there is a tendency to increase the division of poles belonging to the same physical mode into more than one cluster, which represents a disadvantage to the modal identification. If the threshold increased, the tendency decreases, but even when the threshold of 0.10 was used, the number of times that the division happened was higher than when the proposed algorithm was used. In addition, the increase of the threshold value proved to increase the number of global modes with a minimum MAC value below the limit of the stabilization diagram, and decrease these minimum values, which could lead to inaccuracies in the mode shapes' mean. As to the non-identification of some foundation modes, both algorithms performed in the same manner. However, considering that the objective of this paper is the correct identification of the rotor's modes, the identification of a spurious global mode near the first rotor's

Frequency, along with the other findings, demonstrated the superiority of the proposed algorithm's performance.

3.2. Test Rig with Rolling Bearings

To verify the robustness of the proposed algorithm, a distinct system will be analyzed. All data presented in Table 4 will be verified and the results will be briefly presented here, with focus on the identification of the rotor's modes.

For the construction of the stabilization diagrams, the same stabilization and damping ratio limits and stabilization diagram parameters were considered throughout the results showed in this section. Figure 9 displays the stabilization diagram of Test 1 as an example. When compared to the one of Figure 3, this stabilization diagram shows fewer well-defined alignments of stable poles and more poles classified as not stable. However, it is also possible to identify in Figure 9 two well-defined alignments of stable poles near the rotor's modes (Table 5), which, unlike the stabilization diagram of Figure 3, have modal parameters that make them adequate to represent both backward and forward frequencies. These particularities characterize this data set as a source of information about the modal parameters of closely spaced modes and as a real challenge to the identification of the foundation's modes.

After building all stabilization diagrams, the algorithm follows by considering the threshold of 0.01 for the hierarchical clustering of all data sets, and the same one is used in the analyses from the previous section, demonstrating again how easy it is to select this threshold. The additional step was also considered to generate the results of the test rig supported by rolling bearings. The results for the rotor's modes are displayed in Table 10, from which one can see that, even with unfavorable excitation conditions, the algorithm

can extract representative global modes for the rotor, with low standard deviations and modal parameters close to the ones estimated by EMA.

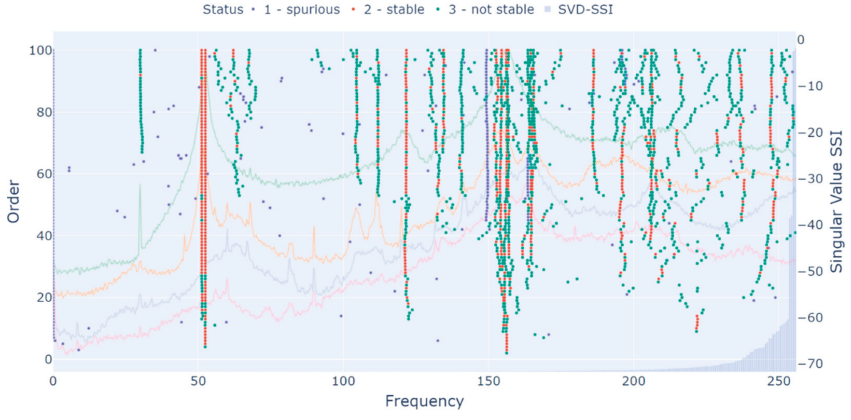


Figure 9. Test 1 (white noise—medium intensity) of the test rig with rolling bearings stabilization diagram.

Table 10. Rotor's global modes for the test rig with rolling bearings.

Mode	Backward						Forward					
	f [Hz]			ζ [%]			f [Hz]			ζ [%]		
Test	**Mean**	**Std.**	**Error**	**Mean**	**Std.**	**Error**	**Mean**	**Std.**	**Error**	**Mean**	**Std.**	**Error**
EMA	51.35	-	-	1.168	-	-	52.65	-	-	0.864	-	-
1	51.12	0.03	0.45%	1.628	0.017	39.38%	52.43	0.00	0.42%	0.682	0.006	21.06%
2	52.07	0.12	1.40%	1.362	0.201	16.61%	53.55	0.02	1.71%	1.027	0.048	18.87%
3	51.04	0.04	0.60%	1.346	0.087	15.24%	52.62	0.01	0.06%	0.607	0.014	29.75%
4	51.06	0.03	0.56%	1.573	0.065	34.67%	52.51	0.04	0.27%	0.733	0.036	15.16%

Comparing Tables 9 and 10, one can observe that the estimation's errors are really close to each other, demonstrating the algorithms' robustness to different datasets.

As mentioned in the previous section, it is expected that the forward and backward frequencies present different mode shapes. Since the test rig supported by rolling bearings provided good results for both frequencies, their mode shapes were compared. Test 1 of Table 4 was once more taken as an example and the MAC value was computed between the mode shapes of all poles from the backward frequency and the mode shapes of all poles from the forward frequencies, producing a MAC matrix of 75×68 (the number of poles from the clusters of the backward and forward frequencies, respectively). The mean, maximum, and minimum MAC values of the matrix were 75%, 82%, and 67%, confirming the expected difference.

Moreover, in order to evaluate the ability of the proposed algorithm to extract the foundation's modes when a different system is considered, Figure 10 displays the extracted modes as black dots, the rotor's modes as continuous lines, and the foundation's modes as dashed lines. From Figure 10, one can see that the algorithm was able to extract several of the foundation's modes from the data of Test 1. When data from different tests are employed, only a few foundation's modes are identified, which could be associated to unfavorable test conditions, and some modes out of the investigated frequency range (80 Hz to 320 Hz) appear. Moreover, the algorithm identifies some extra modes near the foundation mode of 157.4 Hz when data from Tests 1 and 2 are employed. Investigations performed with the same test rig by [25] detected a mode associated to the bearings housing near the frequency of 155 Hz, which would explain these extra identified modes. Therefore, the proposed algorithm demonstrated a good ability to identify the foundation's modes.

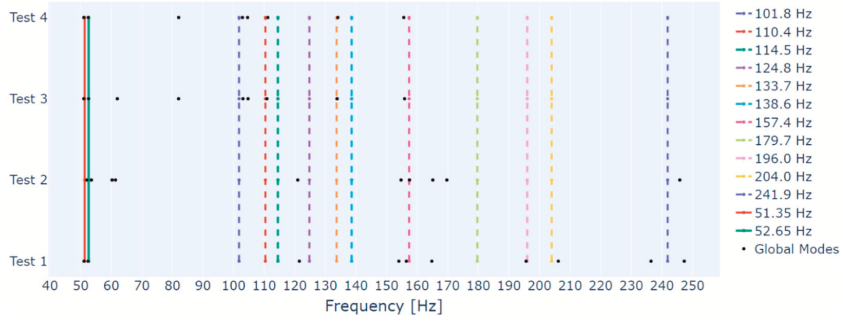


Figure 10. Foundation's modes identified through the proposed algorithm for the test rig with rolling bearings.

As already mentioned, the results of this section were generated considering the additional step; however, the algorithm considering only the essential steps would also be capable of identifying accurate frequencies and damping ratios of all rotors' modes, which was also observed in the results from the test rig supported by hydrodynamic bearings. Hence, the additional step is recommended when a higher precision in the mode shapes estimation is required or when a MAC criterion inside each cluster needs to be respected.

4. Conclusions

In this paper, a new automated algorithm to carry out modal parameter identification on rotating machinery through OMA is proposed. The novelty of the work is precisely the fact that it was developed for the identification of the rotor's modes, and tested for unideal operating conditions that are usually present in the operation of rotating machines. The algorithm was applied through two datasets: vibration signals from a test rig with a rotor supported by hydrodynamic bearings and vibration signals from a test rig with a rotor supported by rolling bearings. Each step of the algorithm was presented, explained, and illustrated, highlighting the differences to other algorithms proposed in the literature, which were mainly developed to deal with signals from structures rather than from rotating machines.

The test in which the operating rotor supported by hydrodynamic bearings is excited by the white gaussian noise of medium intensity was used to illustrate each step of the algorithm. From the results, it was possible to verify that some of the measures proposed by other papers to differentiate physical poles from mathematical and spurious poles are inadequate when the system under analysis is a rotating machine. The results of this data set and of the data sets with other excitation conditions also demonstrated that the proposed algorithm can extract from the stabilization diagram representative and accurate frequencies and damping ratios for both the rotor's and the foundation's modes, even when unfavorable test conditions are present.

Moreover, investigations were carried out to evaluate the performance of the algorithm when the additional step is implemented to the group, with hierarchical clustering and poles with high MAC values within each global mode. From the test with white gaussian noise of medium intensity excitation, the results showed that the additional step can find sets of poles with mode shapes of high correlation. The additional step was also compared with an algorithm that considers a single hierarchical clustering with similarity measure comprising both the frequency difference and the MAC value, as proposed by some previous authors. The results showed that the algorithm proposed in this paper, considering the additional step, presented better results than previous algorithms, especially when the correct identification of the rotor's modes is considered.

When applied to a different system (a rotor supported by rolling bearing), the algorithm was also able to extract from the stabilization diagram representative and accurate

frequencies and damping ratios for both the rotor's and the foundation's modes. These results demonstrated that the proposed algorithm maintained its robustness even when a different system was employed. In addition, the backward and forward frequencies of the first rotor's mode were identified and the mode shapes extracted for each one confirmed that some difference between them is expected.

Therefore, the proposed algorithm proved to be an adequate and promising tool to extract modal parameters of rotating machines in operation. Further investigations are required to improve the extraction of representative mode shapes and the differentiation of the rotor's backward and forward frequencies.

The results were obtained by applying the proposed algorithm to data from test rigs. However, it is expected that it also works on more complex systems. The aim of the ongoing works is to test it in more complex systems, such as engines and compressors, to identify modes from both the rotor and the foundation. Once the algorithm's robustness to more complex equipment is verified, the goal is to use it to monitor the modal parameters of the system and identify failures, given that variations in the modal parameters may be caused by them. With that, one can enable the SHM via OMA.

Author Contributions: Conceptualization, N.R.D. and T.H.M.; methodology, N.R.D., G.C.S. and T.H.M.; software, N.R.D.; validation, N.R.D., G.C.S. and T.H.M.; formal analysis, N.R.D., G.C.S. and T.H.M.; investigation, N.R.D., G.C.S. and T.H.M.; resources, T.H.M.; data curation, N.R.D.; writing—original draft preparation, N.R.D.; writing—review and editing, G.C.S. and T.H.M.; visualization, N.R.D.; supervision, T.H.M.; project administration, T.H.M.; funding acquisition, T.H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Laboratory of Rotating Machinery at Unicamp for the infrastructure support to this research.

Conflicts of Interest: The authors declare no conflict of interest.

List of Abbreviations

Abbreviation	Description
CVA	Canonical Variate Algorithm
EMA	Experimental Modal Analysis
FDE	Frequency Domain Editing
FRF	Frequency Response Function
HVC	Hard Validation Criteria
IOMAC	International Operational Modal Analysis Conference
LSCE	Least Square Complex Exponential
MAC	Modal Assurance Criterion
MPC	Mode Phase Collinearity
MPD	Mode Phase Deviation
ODD	Order Domain Deletion
OMA	Operational Modal Analysis
PCA	Principal Component Analysis
SHM	Structural Health Monitoring
SSI	Stochastic Subspace Identification
SSI-COV	Covariance Driven SSI
SSI-DATA	Data Driven SSI
SVD	Singular Value Decomposition
UPC	Unweighted Principal Components

List of Variables

Variable	Description	Unit
f_s	Sampling Frequency	Hz
f	Natural Frequency	Hz
ζ	Damping Ratio	-
φ	Mode Shape	-
Q_1	First Quartile	Engineering Unit
Q_3	Third Quartile	Engineering Unit
IQR	Difference between the upper and lower quartiles	Engineering Unit

References

1. Farrar, C.R.; Worden, K. An Introduction to Structural Health Monitoring. New Trends in Vibration Based Structural Health Monitoring. In *CISM International Centre for Mechanical Sciences*; Springer: Vienna, Austria, 2010; Volume 520, pp. 1–17. [CrossRef]
2. Lynch, J.P.; Farrar, C.R.; Michaels, J.E. Structural Health Monitoring: Technological Advances to Practical Implementations [scanning the issue]. *Proc. IEEE* **2016**, *104*, 1508–1512. [CrossRef]
3. Rainieri, C.; Fabbrocino, G. Influence of model order and number of block rows on accuracy and precision of modal parameter estimates in stochastic subspace identification. *Int. J. Lifecycle Perform. Eng.* **2014**, *1*, 317–334. [CrossRef]
4. Reynders, E.; De Roeck, G. Reference-based combined deterministic–stochastic subspace identification for experimental and operational modal analysis. *Mech. Syst. Signal Process.* **2008**, *22*, 617–637. [CrossRef]
5. Magalhães, F.; Cunha, A.; Caetano, E. Online automatic identification of the modal parameters of a long span arch bridge. *Mech. Syst. Signal Process.* **2009**, *23*, 316–329. [CrossRef]
6. Reynders, E.; Houbrechts, J.; De Roeck, G. Fully automated (operational) modal analysis. *Mech. Syst. Signal Process.* **2012**, *29*, 228–250. [CrossRef]
7. Neu, E.; Janser, F.; Khatibi, A.A.; Orifici, A.C. Fully Automated Operational Modal Analysis using multi-stage clustering. *Mech. Syst. Signal Process.* **2017**, *84*, 308–323. [CrossRef]
8. Cardoso, R.A.; Cury, A.; Barbosa, F. A clustering-based strategy for automated structural modal Identification. *Struct. Health Monit.* **2018**, *17*, 201–217. [CrossRef]
9. Fan, G.; Li, J.; Hao, H. Improved automated operational modal identification of structures based on clustering. *Struct. Control Health Monit.* **2019**, *26*, e2450. [CrossRef]
10. Wu, G.; He, M.; Liang, P.; Ye, C.; Xu, Y. Automated Modal Identification Based on Improved Clustering Method. *Math. Probl. Eng.* **2020**, *2020*, 16. [CrossRef]
11. Mugnaini, V.; Fragonara, L.Z.; Civera, M. A machine learning approach for automatic operational modal analysis. *Mech. Syst. Signal Process.* **2022**, *170*, 108813. [CrossRef]
12. Amer, M.; Wallaschek, J.; Seume, J.R.; Ventura, C.E. Comparison of different OMA techniques and their application to an axial compressor test rig. In Proceedings of the International Operational Modal Analysis Conference, Vancouver, BC, Canada, 3–6 July 2022.
13. Priou, J.; Gres, S.; Perrault, M.; Guerineau, L.; Döhler, M. Automated uncertainty-based extraction of modal parameters from stabilization diagrams. In Proceedings of the International Operational Modal Analysis Conference, Vancouver, BC, Canada, 3–6 July 2022.
14. Dreher, R.D.; Storti, G.C.; Machado, T.H. Evaluation of an automatic OMA identification method on rotating machinery. In Proceedings of the International Operational Modal Analysis Conference, Vancouver, BC, Canada, 3–6 July 2022.
15. Brandt, A. A signal processing framework for operational modal analysis in time and frequency domain. *Mech. Syst. Signal Process.* **2019**, *115*, 380–393. [CrossRef]
16. Gres, S.; Andersen, P.; Hoen, C.; Damkilde, L. *Orthogonal Projection-Based Harmonic Signal Removal for Operational Modal Analysis, Structural Health Monitoring, Photogrammetry & DIC*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 6, pp. 9–21. [CrossRef]
17. Gres, S.; Andersen, P.; Damkilde, L. *Operational Modal Analysis of Rotating Machinery, Rotating Machinery, Vibro-Acoustics & Laser Vibrometry*; Springer: Cham, Switzerland, 2019; Volume 7, pp. 67–75. [CrossRef]
18. Gres, S.; Döhler, M.; Andersen, P.; Mevel, L. Kalman filter-based subspace identification for operational modal analysis under unmeasured periodic excitation. *Mech. Syst. Signal Process.* **2021**, *146*, 106996. [CrossRef]
19. Gioia, N.; Daems, P.J.; Peeters, C.; El-Kafafy, M.; Guillaume, P.; Helsen, J. *Influence of the Harmonics on the Modal Behavior of Wind Turbine Drivetrains, Rotating Machinery, Vibro-Acoustics & Laser Vibrometry*; Springer: Cham, Switzerland, 2019; Volume 7, pp. 231–238. [CrossRef]
20. Peeters, C.; Gioia, N.; Helsen, J.; Guillaume, P. Identification of Noise Vibration and Harshness Behavior of Wind Turbine Drivetrain under different operating conditions. *Energies* **2019**, *12*, 3401. [CrossRef]
21. Dreher, R.D.; Storti, G.C.; Machado, T.H. Directional coordinates for the identification of backward and forward frequencies of rotating machines via OMA. In Proceedings of the International Operational Modal Analysis Conference, Vancouver, BC, Canada, 3–6 July 2022.

22. Zivanovic, M.; Plaza, A.; Iriarte, X.; Carlosena, A. Harmonic removal for wind turbines. In Proceedings of the International Operational Modal Analysis Conference, Vancouver, BC, Canada, 3–6 July 2022.
23. Peeters, B.; De Roeck, G. Reference-based Stochastic Subspace Identification for Output-only Modal Analysis. *Mech. Syst. Signal Process.* **1999**, *13*, 855–878. [CrossRef]
24. Peeters, B.; De Roeck, G. Stochastic System Identification for Operational Modal Analysis: A Review. *ASME J. Dyn. Syst. Meas. Control.* **2001**, *123*, 659–667. [CrossRef]
25. Storti, G.; Machado, T. The use of operational modal analysis in the process of modal parameters identification in a rotating machine supported by roller bearings. *J. Mech. Sci. Technol.* **2021**, *35*, 471–480. [CrossRef]
26. Juang, J.N.; Pappa, R. An Eigensystem Realization Algorithm for Modal Parameter Identification and Model Reduction. *J. Guid. Control. Dyn.* **1985**, *8*, 620–627. [CrossRef]

sensors

MDPI

Article

Rolling Bearing Performance Degradation Assessment with Adaptive Sensitive Feature Selection and Multi-Strategy Optimized SVDD

Zhengjiang Feng [1,2], Zhihai Wang [1,2,*], Xiaoqin Liu [1,2] and Jiahui Li [1,2]

[1] Key Laboratory of Advanced Equipment Intelligent Manufacturing Technology of Yunnan Province, Kunming University of Science & Technology, Kunming 650500, China
[2] Faculty of Mechanical & Electrical Engineering, Kunming University of Science & Technology, Kunming 650500, China
* Correspondence: wzh_kust@163.com; Tel.: +86-136-6872-5353

Abstract: In light of the problems of a single vibration feature containing limited information on the degradation of rolling bearings, the redundant information in high-dimensional feature sets inaccurately reflecting the reliability of rolling bearings in service, and assessments of the degradation performance being disturbed by outliers and false fluctuations in the signal, this study proposes a method of assessing rolling bearings' performance in terms of degradation using adaptive sensitive feature selection and multi-strategy optimized support vector data description (SVDD). First, a high-dimensional feature set of vibration signals from rolling bearings was extracted. Second, a method combining the Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) and K-medoids was used to comprehensively evaluate the features with multiple evaluation indicators and to adaptively select better degradation features to construct the sensitive feature set. Next, multi-strategy optimization of the SVDD model was carried out by introducing the autocorrelation kernel regression (AAKR) and a multi-kernel function to improve the ability of the evaluation model to overcome outliers and false fluctuations. Through validation, it could be seen that the method in this study uses samples of rolling bearings in the healthy early stage to establish the evaluation model, which can adaptively determine the starting point of the bearing's degradation. The stability and accuracy of the model were effectively improved.

Keywords: performance degradation assessment; rolling bearing; SVDD; feature selection; multi-strategy optimization

Citation: Feng, Z.; Wang, Z.; Liu, X.; Li, J. Rolling Bearing Performance Degradation Assessment with Adaptive Sensitive Feature Selection and Multi-Strategy Optimized SVDD. *Sensors* **2023**, *23*, 1110. https://doi.org/10.3390/s23031110

Academic Editor: Jongmyon Kim

Received: 22 November 2022
Revised: 5 January 2023
Accepted: 10 January 2023
Published: 18 January 2023

1. Introduction

The failure of rolling bearings, as one of the key components of rotating machinery, leads to the breakdown of the whole mechanical system [1,2]. During the in-service period, the performance of rolling bearings degrades irreversibly due to fatigue, wear, and other reasons. Effective assessment in the performance degradation assessment (PDA) of rolling bearings in the service phase can help organize maintenance in a targeted manner to prevent failure from occurring and improve the operational reliability of the whole machine.

Assessment of the degradation in the performance of rolling bearings mainly includes three steps: acquisition of the rolling bearings' monitoring data, feature extraction, and establishment of the model for assessing the degradation. The degradation mechanism of rolling bearings is complex, and the vibration signals of rolling bearings are nonlinear and nonstationary. A single feature contains less information about bearings degradation and has poor anti-interference ability, so it cannot accurately characterize the whole process of degradation during performance. Constructing a high-dimensional feature set can comprehensively reflect the information on the bearings' degradation and benefit from the complementarity of the differences among features, but some features are unrelated to

degradation. Irrelevant features can be eliminated using feature selection methods, which can be classified into two categories based on whether the methods are independent of subsequent learning algorithms, namely filters and wrappers [3]. The wrapper method is computationally complex and less versatile because it requires several iterations in combination with subsequent learning algorithms to find the best combination of features. Feature evaluation [4,5] is one of the commonly used filtering methods and is independent of the subsequent learning algorithm. It can quickly remove irrelevant features with high generality and interpretability, so it is often used in engineering applications [6]. Although feature ranking can be achieved using feature evaluation, the selection of the feature set relies heavily on experts' prior knowledge, which reduces the efficiency of the algorithm and may introduce subjective errors.

In general, data-driven fault diagnosis techniques use machine learning algorithms to identify fault status to train predictive models based on the condition data collected under normal and different faulty states [7]. The training of the model is based on the condition data collected in the health stage and the degradation stage. Most of these data-driven approaches rely entirely on data at different stages [8,9]. For assessing the degradation in the performance of rolling bearings, the process of bearing degradation usually consists of the healthy phase and the degradation phase. In actual production, fewer data are available on the degradation stage, and sometimes, only data from the healthy state are available. Knowledge-based methods often need to capture a large number of samples in advance to identify faults [10]. Rai et al. [11] performed K-medoids clustering to train a model using the full-life feature set of bearings obtained with empirical modal decomposition (EMD) and calculated the dissimilarity between the bearing samples to be tested and the clustering centers used as health indicators. Pan et al. [12] used lifting wavelet packet decomposition and fuzzy c-means combined with the affiliation function to characterize the severity of bearings failure. Adaptive determination of the start time of the degradation phase (first predicting time, FPT) can effectively trigger an early warning to carry out condition-based maintenance. Heng et al. [13] used principal component analysis (PCA) to fuse the time domain features to extract the life cycle health index of rolling bearings and then divided the performance stages according to the amplitude of the change trends of vibration signals. These methods are suitable for obtaining health indicators but ignore the difficulties of obtaining data from the degradation stage and carrying out secondary determination of FPT. Finally, the data monitoring process inevitably suffers from the interference of noise and environmental changes, which lead to outliers and false fluctuations in the data. Liu et al. [14] used the features extracted from the time domain combined with the SVDD model to monitor the faults in rolling bearings and to overcome the interference of random fluctuations by using a decision strategy. The authors of [15] used the method of repairing the evaluation results, which created problems such as subjectivity and reduced interpretability. The recognition ability and robustness of models in mechanical learning are always required [16,17]. Therefore, there are still shortcomings in using sensitive feature sets for a PDA of bearings, such as the models' reliance on data for the full life cycle of the bearings, FPT needing to be determined twice, and the evaluation model being easily affected by outliers and false fluctuations. How to achieve efficient fault diagnosis using only health data has attracted our attention.

In summary, extracting effective feature sets is a prerequisite for accurately assessing the performance of bearings, and improving the ability of the model to overcome outliers and false fluctuations is one of the critical tasks in assessing degradation. Accordingly, a rolling bearing performance degradation assessment method with the combination of adaptive sensitive feature selection and multi-strategy optimized SVDD was proposed in this paper. The specific contributions of this study are described below. TOPSIS-Kmedoids, an adaptive sensitive feature selection method, was proposed, which could determine the adaptive sensitive feature set without prior knowledge. In addition, SVDD was optimized using a multi-strategy, in which AAKR was introduced to correct the errors in monitoring data, and a multi-kernel function was constructed to improve the learning ability and

generalization ability of the model. Lastly, the effectiveness of the proposed method was verified using the XJTU-SY dataset for the full life cycle of rolling bearings from Xi'an Jiaotong University, the PHM2012 Data Challenge dataset for the full life cycle of rolling bearings, and a set of data from a self-made bench test of accelerated fatigue in rolling bearing.

2. Determination of the Adaptive Sensitive Feature Set

2.1. Feature Extraction

In the field of prognostics and health management (PHM) of rolling bearings, the vibration signal is one of the most commonly used means because it contains much information on degradation. Feature extraction can reveal information on the performance of sensor data. Twenty-four commonly used statistical features of vibration were extracted from the time domain and the frequency domain of vibration signals, as shown in Table 1, where F_1–F_7 are the frequency domain features, s_i is the amplitude of the vibration data, sk_i is the spectral amplitude of the vibration data, and f_i is the frequency of the vibration data. For the data from two accelerometers, the features listed in Table 1 were extracted separately to form the high-dimensional set, where n is the feature length and m is the number of features.

Table 1. Time domain and frequency domain features.

Name	Equation	Name	Equation
Mean value	$\mu_s = \frac{1}{n}(\sum_{i=1}^{n} s_i)$	Standard deviation	$\sigma_s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i - \mu_S)^2}$
Average amplitude	$\mu_a = \frac{1}{n}\sum_{i=1}^{n}\|s_i\|$	Variance	$\sigma_s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(s_i - \mu_S)^2$
Maximum value	$F_{\max} = \max(s_i)$	Kurtosis	$F_{kurt} = \frac{1}{n}\sum_{i=1}^{n}(s_i - \mu_S)^4$
Minimum value	$F_{\min} = \min(s_i)$	Skewness	$F_{SK} = \frac{1}{n}\sum_{i=1}^{n}(s_i - \mu_S)^3$
Peak value	$F_{\|\max\|} = \max(\|s_i\|)$	Waveform index	$F_{WI} = F_{RMS}/\mu_a$
Peak to peak value	$F_{P2P} = F_{\max} - F_{\min}$	Peak index	$F_{PI} = F_{\|\max\|}/F_{RMS}$
Root mean square	$F_{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {s_i}^2}$	Impulse index	$F_{IF} = F_{\|\max\|}/\mu_a$
Root amplitude	$F_{RA} = (\frac{1}{n}\sum_{i=1}^{n}\sqrt{\|s_i\|})^2$	Tolerance index	$F_{MF} = F_{\|\max\|}/F_{RA}$
Kurtosis index	$F_{KI} = \frac{\sum_{i=1}^{n}(s_i-\mu_S)^4}{(n-1){\sigma_s}^4}$	F_1	$\mu_{sk} = \frac{1}{n}\sum_{i=1}^{n} sk_i$
F_2	$\sigma_{sk} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(sk_i - \mu_{Sk})^2}$	F_3	$F_3 = \sqrt{\frac{\sum_{i=1}^{n}(sk_i-\mu_{Sk})^3}{n\sqrt{{\sigma_{sk}}^3}}}$
F_4	$F_4 = \sqrt{\frac{\sum_{i=1}^{n}(sk_i-\mu_{Sk})^4}{n{\sigma_{sk}}^2}}$	F_5	$F_5 = \frac{1}{n}\sum_{i=1}^{n} f_i$
F_6	$F_6 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - F_{22})^2 sk_i}$	F_7	$F_7 = \sqrt{\frac{\sum_{i=1}^{n} {f_i}^2 sk_i}{\sum_{i=1}^{n} sk_{ii}}}$

2.2. Feature Evaluation with Multiple Criteria

The quality of features significantly affects the results when assessing the degradation. Good features should correlate strongly with the bearings' degradation process with monotonic increasing or decreasing characteristics and robustness to outliers [3]. Most of the existing methods for evaluating the quality of features use a single-indicator evaluation scheme. To select excellent features, this study simultaneously considered the three indicators of monotonicity, correlation, and robustness [18], which are described as follows:

(1) Monotonicity: The degradation of rolling bearings' performance is an irreversible process, so the features should be able to monotonically characterize the process of rolling bearings from operation to failure. Because the time vector $\mathbf{t}_i$ is strictly monotonic, the correlation coefficient $Mon(\mathbf{A}_i)$ between the feature vector $\mathbf{A}_i$ and the time vector $\mathbf{t}_i$ is used to measure the monotonicity of the feature [19]. In practice, rolling bearings often show a nonlinear degradation trend. The Spearman's rank correlation coefficient is widely applicable and is more sensitive to nonlinear correlation [1], so

this study calculated the Spearman's rank correlation coefficient as the monotonicity index of the feature. The monotonicity score's equation is shown in Equation (1) [3] as:

$$Mon(\mathbf{A}_i) = 1 - \frac{6 \times \sum_{i=1}^{n} [rank(\mathbf{A}_i) - rank(\mathbf{t}_i)]^2}{n(n^2-1)} \quad (1)$$

where $rank(\mathbf{A}_i)$ and $rank(\mathbf{t}_i)$ indicate $\mathbf{A}_i$ and $\mathbf{t}_i$ in ascending order, respectively, and n is the feature length.

(2) Robustness: During the use of rolling bearings, the signal acquisition process is inevitably disturbed by the environment, changes in the working condition, and noise. The robustness index $Rob(\mathbf{A}_i)$ is used to measure the tolerance of the features to random noise and abnormal values [19]. Equation (2) is used for calculating the robustness score of features, which is a widely used and interpretable equation for calculating the robustness index of features [1], as follows:

$$Rob(\mathbf{A}_i) = \frac{\sum_{i=1}^{n} \exp(-|\mathbf{f}_i^r / \mathbf{A}_i|)}{n} \quad (2)$$

where $\mathbf{f}_i^t$ and $\mathbf{f}_i^r$ are the trend and residual values of the ith feature $\mathbf{A}_i$, respectively. These two items can be obtained using smoothing methods and satisfying the equation $\mathbf{A}_i = \mathbf{f}_i^t + \mathbf{f}_i^r$ [3].

(3) Correlation: The correlation index $Cor(\mathbf{A}_i)$ is used to measure whether the feature can capture the trend of the degradation in performance across the life cycle of rolling bearings [20]. The equation for calculating the correlation score is shown in Equation (3), which can measure the change trend of the features across the whole life cycle [3], as:

$$Cor(\mathbf{A}_i) = \frac{\left| n \sum_{i=1}^{n} (i\mathbf{f}_i^t) - \sum_{i=1}^{n} \mathbf{f}_i^t \sum_{i=1}^{n} i \right|}{\sqrt{[n \sum_{i=1}^{n} (\mathbf{f}_i^t)^2 - (\sum_{i=1}^{n} \mathbf{f}_i^t)^2] - [n \sum_{i=1}^{n} i^2 - (\sum_{i=1}^{n} i)^2]}} \quad (3)$$

where $\mathbf{f}_i^t$ is the trend values of the ith feature $\mathbf{A}_i$.

All the indexes above are positive indicators; that is, the higher the evaluation score, the better the feature's quality.

2.3. Adaptive Sensitive Feature Selection

A single metric can barely make a comprehensive and accurate evaluation of the degradation features. Sensitive features should be selected using integration of the evaluation indicators mentioned in Section 2.2. Linear weighting is commonly used to construct the comprehensive metrics when relying on multiple metrics to evaluate features, and the allocation of weights will directly impact the results of evaluation. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method is a commonly used comprehensive evaluation method that constructs comprehensive metrics without relying on prior knowledge to subjectively determine weights [3]. Meanwhile, the selection of sensitive feature sets has the disadvantage of relying on the prior knowledge of experts. The K-medoids algorithm is a robust clustering algorithm, which can divide features into clusters according to rules to realize adaptive classification of the features. Therefore, the TOPSIS–K-medoids method was applied for a comprehensive evaluation of the features with the evaluation indexes constructed using Equations (1)–(3) and for constructing the adaptive sensitive feature set. The key steps are shown in Figure 1 and described below.

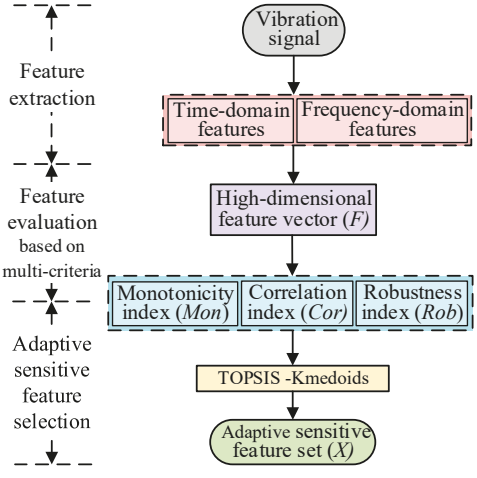


Figure 1. Process for selecting the adaptive sensitive features.

Step 1: Construction and normalization of the evaluation matrix. The feature evaluation matrix $\mathbf{Q} = [Mon(\mathbf{A}_i); Cor(\mathbf{A_i}); Rob(\mathbf{A_i})]$ is constructed using Equations (1)–(3), where $M \times N$ is a matrix of $\mathbf{Q}$ dimensions, M is the number of features (M = 48 in this study), and N is the number of evaluation metrics (N = 3 in this study). The evaluation matrix is normalized using Equation (4):

$$y_{ij} = q_{ij} / \sqrt{\sum_{i=1}^{M} {q_{ij}}^2} \tag{4}$$

where $q_{ij} (1 \le i \le M, 1 \le j \le N)$ represents the elements in the evaluation matrix $\mathbf{Q}$ and denotes the standardized value of the jth evaluation metric for the ith feature.

Step 2: Calculation of the TOPSIS score. The maximum and minimum values of each evaluation index are obtained and defined as the superior solutions Y_j^+ and inferior solutions Y_j^-, then the TOPSIS scores S_i of the features are calculated according to Equation (5) [4]:

$$S_i = \sqrt{\sum_{i=1}^{M} \left(Y_j^- - y_{ij}\right)^2} / [\sqrt{\sum_{i=1}^{M} \left(Y_j^+ - y_{ij}\right)^2} + \sqrt{\sum_{i=1}^{M} \left(Y_j^- - y_{ij}\right)^2}] \tag{5}$$

where the TOPSIS score is positively correlated with the signal, such that the higher the score S_i, the richer the degradation information contained in the feature.

Step 3: Feature clustering. The K-medoids algorithm was improved from the K-means method. K-medoids is a clustering algorithm with good robustness. To achieve adaptive classification of the features, K-medoids is used to divide the data into class clusters according to certain rules, so that samples of the class cluster are similar. In order to adaptively determine the sensitive features set, the feature selection process is transformed into the K-medoids clustering problem with the TOPSIS score. The core idea of the K-medoids algorithm is to divide the feature scores, as obtained in Step 2, into clusters under the condition that the sum of dissimilarities between the cluster's elements and the cluster's center is minimized [6]. The highest cluster is then extracted and determined to be the sensitive feature set. The sum of algorithmic dissimilarities J is calculated as follows:

$$\min J = \sum_{j=1}^{k} \sum_{x_i = c_j} D(S_i, o_j) \tag{6}$$

where c_j is the jth cluster, o_j is the jth medoid, and $D(S_i, o_j)$ is the distance between S_i and o_j. To overcome the problem that the K-medoids clustering algorithm can easily fall into the local optimal state because of improper initial point selection, selecting samples with the relative distance as the initial clustering center can effectively improve this situation. A trade-off between similarity and the weighted Euclidean distance can improve the classification's accuracy [20]. Therefore, the improved K-medoids clustering algorithm selects the initial clustering center with a more considerable distance and divides the features into clusters according to the similarity of the weighted Euclidean distances. The weight is the proportion of the feature's score to the sum of all features' scores, and the weighted distance calculation equation is:

$$D(S_i, o_j) = \| S_i - o_j \|^2 \left(S_i / \sum_{i=1}^{n} S_i \right) \tag{7}$$

Step 4: Adaptive sensitive feature set. The medoid is adjusted using iteration according to Equation (6) until the center point no longer changes. The cluster with a large TOPSIS value of M in the center of the cluster is determined as the adaptive sensitive feature set $\mathbf{X}^{L \times m}$, where L is the length of the sensitive features, m is the number of selected features, and $\mathbf{x}_i$ is the ith sensitive feature.

3. Multi-Strategy Optimized Support Vector Data Description

3.1. Support Vector Data Description

SVDD is an effective one-class classification algorithm proposed by Tax et al. [21] in 1999. The core idea of the SVDD algorithm is as follows: The target samples are first mapped to the high-dimensional feature space using nonlinear transformation, then a minimum hypersphere containing most, if not all, the training samples is established in the feature space. In contrast, the nontarget values are distributed outside the hypersphere as much as possible [13]. SVDD uses the hypersphere as its decision surface, and a schematic diagram of this is shown in Figure 2. The center O and the radius R of the sphere are the decision variables of the hypersphere, the samples on the boundary of the hypersphere are the support vectors, and the samples outside the hypersphere are outliers.

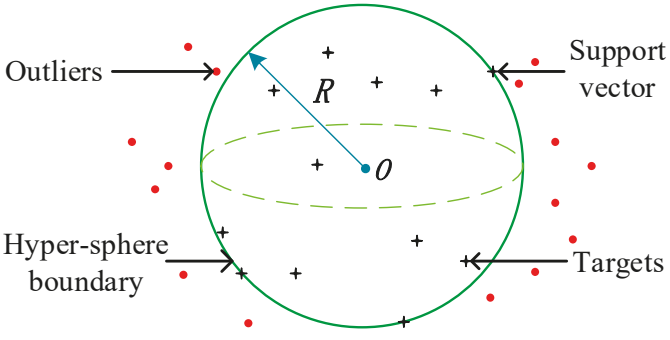


Figure 2. Schematic diagram of SVDD.

SVDD is a semi-supervised model in which only one type of target sample is required for the model's training, that is, the model can be trained with samples from normal stages. The labeled samples are the training set of SVDD, which needs to construct the minimum radius hypersphere. The objective function can be described using Equation (8):

$$F(R, O) = \min R^2 + C \sum_{i=1}^{n} \xi_i \tag{8}$$

The objective function must also satisfy the constraints $\| \mathbf{x}_i - o \| \leq R^2 + \xi_i$, $\xi_i \geq 0$, where $\mathbf{x}_i$ represents the labeled samples; ξ_i is the relaxation variable, which allows a small number of labeled samples to be distributed outside the hypersphere to reduce the effect of outliers on the radius of the hypersphere; and C is the penalty coefficient, which acts to maintain the balance between the size of the radius R and the number of samples falling outside the sphere.

Equation (8) is a quadratic convex optimization problem with univariate variables. By introducing Lagrangian multipliers, constraints are fused into the objective functions to form dual forms, and the following results can be obtained:

$$\max L = \sum_{i=1}^{n} \alpha_i (\mathbf{x}_i \cdot \mathbf{x}_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{9}$$

where $\mathbf{x}_j$ represents the labeled samples, and $(\mathbf{x}_i \cdot \mathbf{x}_j)$ is the inner product of $\mathbf{x}_i$ and $\mathbf{x}_j$.

To cope with the nonlinearity problem, the kernel function is introduced to replace the inner product. The kernel function maps the data to the high-dimensional feature space, which makes the nonlinear data easier to linearly separate in the high-dimensional feature space. According to the Karush–Kuhn–Tucher condition [22], it is known that the training samples satisfying $\alpha_i = 0$ will be wrapped inside the hypersphere; those satisfying $0 < \alpha_i < C$ will be the support vector; and those with $\alpha_i = C$ are judged to be outliers. The radius of hypersphere is obtained as follows:

$$R = \left[K(\mathbf{x}_i \cdot \mathbf{x}_i) - 2 \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \right]^{1/2} \tag{10}$$

where $K(\mathbf{x}_i \cdot \mathbf{x}_i)$ is the kernel function.

The equation for calculating the distance between the new sample $\mathbf{x}_n$ and the center R of the hypersphere is:

$$D = \left[K(\mathbf{x}_n \cdot \mathbf{x}_n) - 2 \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i \cdot \mathbf{x}_n) + \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \right]^{\frac{1}{2}} \tag{11}$$

3.2. Construction of the Multi-Kernel Function

Single-kernel functions have limitations in dealing with outliers and false fluctuations. Different kernel functions have different levels of efficacy, and multiple kernel functions combine different types of kernel functions, which can combine good learning ability and generalization. Using multi-kernel functions can make the results of SVDD more robust [23,24]. The methods of constructing multi-kernel functions include multi-scale kernels and synthetic kernel functions. The synthetic kernel approach has high learning ability and generalization ability and low operational complexity. Therefore, this method was used to construct multi-kernel functions as follows:

$$K_m(\mathbf{x}_i, \mathbf{x}_j) = \omega K_1(\mathbf{x}_i, \mathbf{x}_j) + (1 - \omega) K_2(\mathbf{x}_i, \mathbf{x}_j) \tag{12}$$

where $K_m(\mathbf{x}_i, \mathbf{x}_j)$ represents the multi-kernel functions; ω represents the weights, $0 < \omega < 1$; and $K_1(\mathbf{x}_i, \mathbf{x}_j)$ and $K_2(\mathbf{x}_i, \mathbf{x}_j)$ are single-kernel functions.

The single-kernel functions include Gaussian, Sigmoid, and Laplace kernel functions, and the kernel functions are calculated using:

$$\begin{cases} k_{Gauss}(\mathbf{x}_i, \mathbf{x}_j) = \exp(- \| \mathbf{x}_i - \mathbf{x}_j \|^2 / \sigma_1^2) \\ k_{\mathrm{Tanh}}(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\sigma_2 \mathbf{x}_i{}^T \mathbf{x}_j + \sigma_3) \\ k_{Lapl}(\mathbf{x}_i, \mathbf{x}_j) = \exp(- \| \mathbf{x}_i - \mathbf{x}_j \|^2 / \sigma_4) \end{cases} \tag{13}$$

where σ_1, σ_2, σ_3, and σ_4 are the kernel parameters.

3.3. Auto-Associative Kernel Regression

The monitoring data of rolling bearings cannot avoid outliers and false fluctuations due to the interference of noise, working conditions, and environmental changes. These may greatly impact the performance of the model [25]. Using auto-associative kernel regression to reconstruct the signals can reduce the influence of outliers, so as to improve the robustness and recognition accuracy of the model [26,27]. AAKR was computationally efficient without relying on expert experience to adjust the parameters [3]. Therefore, AAKR was introduced to correct errors in the monitoring data by reconstructing the current feature matrix using the values of health history features to improve the evaluation ability of the models of degradation.

The core idea of AAKR is to map the characteristic matrix at the current moment to estimate the characteristic matrix of rolling bearings in the healthy state. AAKR maps the characteristic matrix at the current moment $\mathbf{X}_t$ in the source space of degradation conditions to the data of the expected state $\overset{\wedge}{\mathbf{X}}_t$ in the target space of normal conditions using:

$$\overset{\wedge}{\mathbf{X}}_t = \sum_{i=1}^{m}(\omega_i \cdot \mathbf{X}^o)/\sum_{i=1}^{m}\omega_i \tag{14}$$

where m is the number of optimal degradation features selected, ω_i represents the weights, and ω_i is determined by the similarity between $\mathbf{X}^o$ and $\mathbf{X}_t$. AAKR uses a Gaussian radial basis function as the kernel for mapping, and the values of ω_i are calculated as follows:

$$\omega_i = e^{-di^2/2h^2}/\sqrt{2\pi h^2} \tag{15}$$

where h is the kernel's bandwidth and di^2 is the distance between $\mathbf{X}^o$ and $\mathbf{X}_t$. Both distance similarity and spatial similarity were considered, and the equation is as follows:

$$di^2 = 1 + (\mathbf{X}_t - \mathbf{X}^o)^T S^{-1}(\mathbf{X}_t - \mathbf{X}^o) - (\mathbf{X}_t \cdot \mathbf{X}^o)/(\left|\mathbf{X}_t\right| \cdot \left|\mathbf{X}^o\right|) \tag{16}$$

where S is the diagonal matrix. The value of the diagonal line is the variance of the historical observation matrix.

3.4. Parameters of SVDD and Indicators of Degradation

The parameters C, σ, and ω of SVDD were determined using particle swarm optimization. The fitness function of the minimum number of support vectors was used for SVDD parameter optimization [14,28], and the calculation equation Fit of fitness function is as follows:

$$Fit = N_{SV}/N_a \tag{17}$$

where N_{SV} is the number of support vectors; and N_a is the total number of training samples. In order to ensure the robustness of SVDD, the minimum number of support vectors is set to 5% of the total number of training samples, and the optimization range of C is set to $[1/N_a, 1/0.05N_a]$. The optimization range of another parameter σ is [0.01, 10] [28]. The optimization range of another parameter ω is [0.01, 1].

To carry out a PDA of rolling bearings, the radius of the hypersphere was determined by carrying out multi-strategy optimized SVDD training with some of the previous normal samples. In the testing stage, the distance D between the test samples and the center of the hypersphere was calculated as the health indicator (HI) according to Equation (11). When $D \leq R$, this indicated that the bearing was in the healthy stage; when $D > R$, this indicated that the bearing had degraded, and a larger value indicated that the degradation of the bearing was more serious. Moreover, when the HI exceeded the threshold value five times in a row, the point where the threshold value was exceeded for the first time was determined as the FPT.

4. Steps of the Algorithm for Assessing Degradation

The process used for a PDA of rolling bearings with an adaptive sensitive feature set and multi-strategy optimized SVDD is shown in Figure 3.

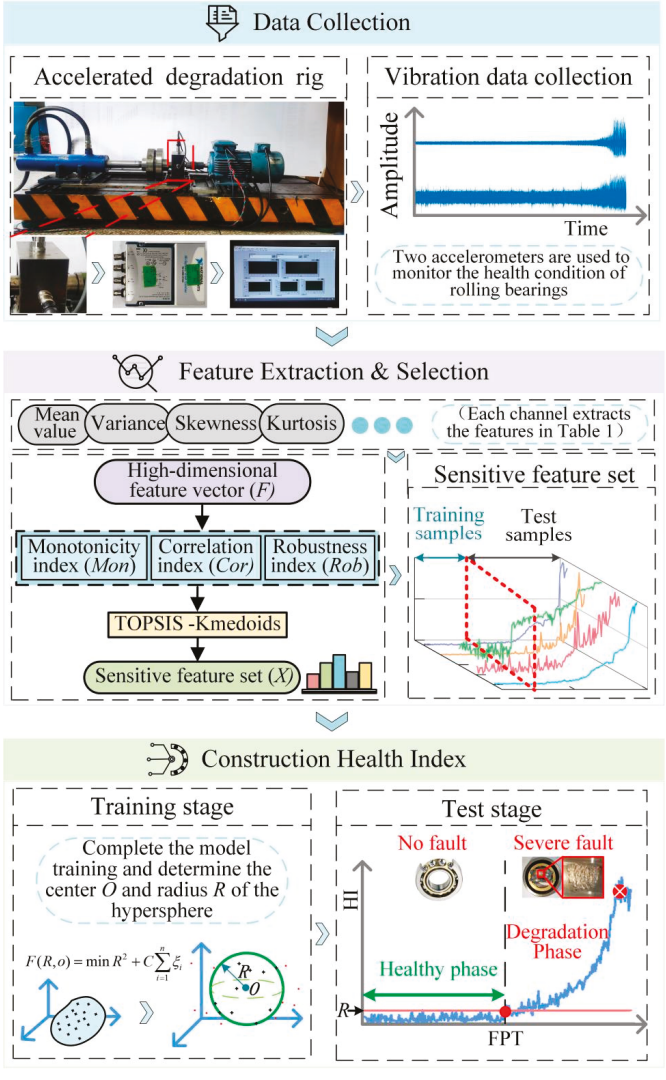


Figure 3. The technical process of the proposed method.

The main steps are as follows:

Step 1: Data acquisition. Obtain vibration signals of the rolling bearing degradation test platform.

Step 2: Construction of the adaptive sensitive feature set. The multi-domain high-dimensional feature set **A** is constructed according to Table 1, and the adaptive sensitive feature set **X** is determined with the TOPSIS–K-medoids method described in Section 2.3.

Step 3: Optimize SVDD with multiple strategies. AAKR is introduced to correct the errors in the monitoring data, and a multi-kernel function is constructed to improve the learning ability and generalization ability of the model.

Step 4: Complete the training of the model. The samples in the early normal state of the sensitive feature set are taken as the training data and are determined to be the historical observation matrix. Training of the SVDD hypersphere using multi-strategy optimization is completed to obtain the hypersphere's radius R and center O.

Step 5: Assessment of the degradation in performance. The test samples are inputted into the completed model, and the performance is evaluated according to the distance value D outputted by the model. Meanwhile, R is set as the adaptive alarm threshold. When $D > R$ occurs several times in a row, the first occurrence point is determined as the FPT, and an early warning is given regarding the degraded state of the rolling bearings.

5. Experimental Verification

5.1. Case 1: XJTU-SY Bearing Datasets

5.1.1. Experimental Description of Case 1

The XJTU-SY rolling bearing dataset was used for verification. Figure 4 shows the platform used for the accelerated bearing degradation test in the experiment, which consisted of the test bearings, an AC motor, a controller of the motor's speed, a hydraulic loading system, and other components [29]. Two accelerometers (model PCB352C33) were used to collect the horizontal and vertical vibration signals of the bearing across the entire life cycle, with a sampling frequency of 25.6 kHz. The signals were sampled for 1.28 s every 1 min during the experiment. The Illustration of sampling parameters is shown in Figure 5.

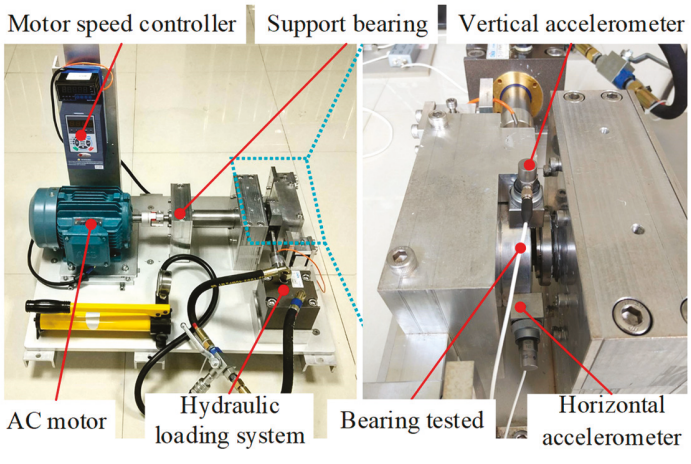


Figure 4. Platform used for the accelerated degradation test of the XJTU-SY bearing datasets.

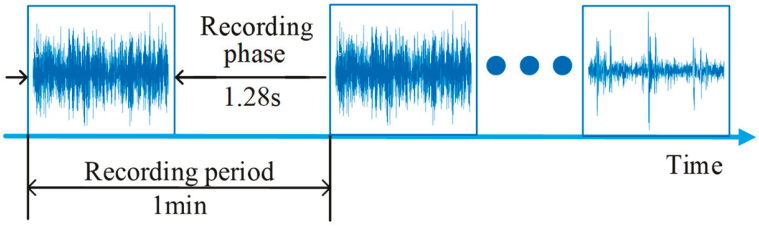


Figure 5. Illustration of the sampling parameters for the vibration signals.

The subset Bearing 1–3 was used for the experiment and analysis, which collected 158 samples. Figure 6 shows the time domain waveform of the data for the bearings' complete life cycle.

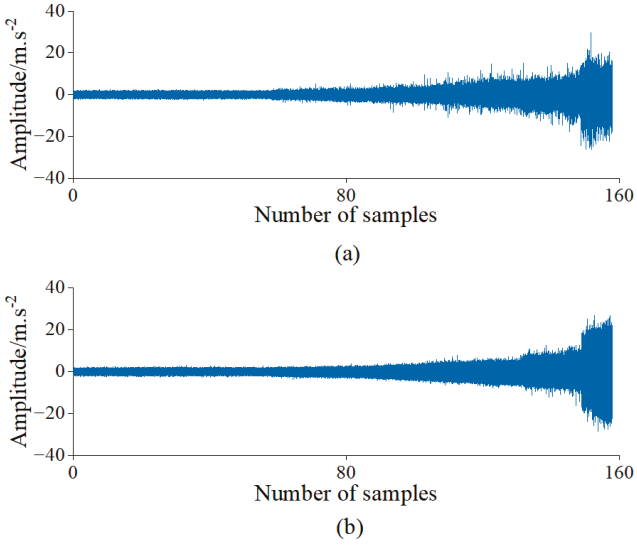


Figure 6. Bidirectional acceleration waveforms of rolling bearings in the time domain of Case 1. (**a**) Horizontal vibration signals; (**b**) Vertical vibration signals.

The PDA of rolling bearings was carried out according to the technical scheme in Figure 3. First, the time domain and frequency domain features of the vibration signal were extracted to construct a high-dimensional feature matrix and normalize it. Next, in order to remove the invalid features, the evaluation index of each feature was calculated, and the adaptive sensitive feature set was constructed according to Equations (4)–(7). The adaptive sensitive features thus determined are shown in Figure 7 (the notation VX-FX represents the corresponding features in Table 1 extracted from the vibration signal in the horizontal direction, and VY-FX represents the corresponding features from the vibration signal in the vertical direction). As the performance of the rolling bearing deteriorates, the sensitive features change according to different patterns. Each characteristic contains different information about the degradation of the rolling bearing. Finally, the samples from the early part of the healthy stage (i.e., the first 25% of all the samples) were selected as the training samples to complete the training of the SVDD model after multi-strategy optimization. Next, the test samples were fed into the model obtained using training, and the variation trend of the distance D from each sample to the center of the sphere was recorded.

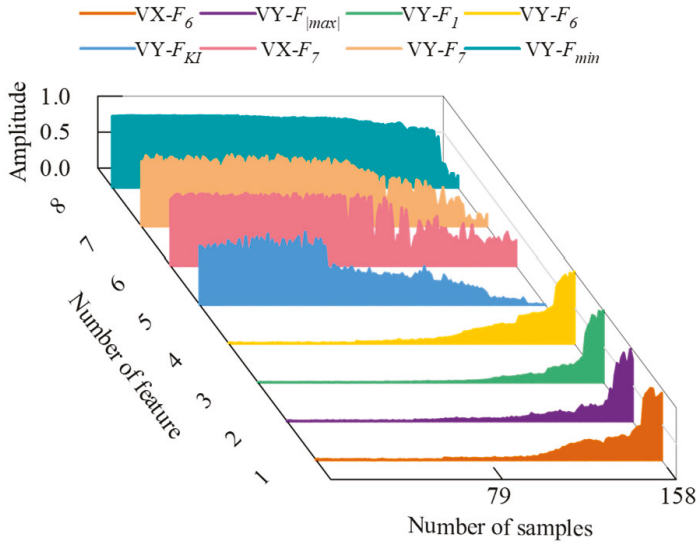


Figure 7. Sensitive features.

5.1.2. Experimental Result of Case 1

To verify the effectiveness of the proposed method, commonly used degradation assessment methods were selected for comparison, including (a) constructing performance metrics by fusing the sensitive features with a PCA, (b) using root mean square (RMS) feature metrics, (c) combining a continuous hidden Markov model (CHMM) with the sensitivity to build assessment metrics, and (d) using the original SVDD combined with sensitive features to build assessment metrics. Method (d) and the proposed method could determine the warning threshold of degradation adaptively, and the rest of the methods determined the threshold using the three principles of the international engineering standard ISO-10816 [30]. When the degradation index exceeded the threshold five consecutive times, the first point where the threshold was exceeded was determined to be the FPT [31].

Figure 8a–e shows the health indicators of each method. In Figure 8e, in the first 58 samples, the HI values were lower than the warning threshold, indicating that the bearing was in a healthy state. From the 59th sample onwards, the HI values exceeded the threshold and gradually increased, indicating that the rolling bearing's performance had started to deteriorate. Envelope spectrum analysis was carried out on Samples 58 and 59, and the analytical results are shown in Figure 9. Compared to Figure 9a, the envelope spectrum in Figure 9b shows the rotation frequency of 32.0 Hz and the outer ring fault has the characteristic frequency of 109.4 Hz and its multiplier, indicating that the rolling bearing was in the healthy state before Sample 58. This shows that the proposed method accurately determined the FPT, while Method (c) determined a wrong FPT, and Method (a) determined the FPT obviously later.

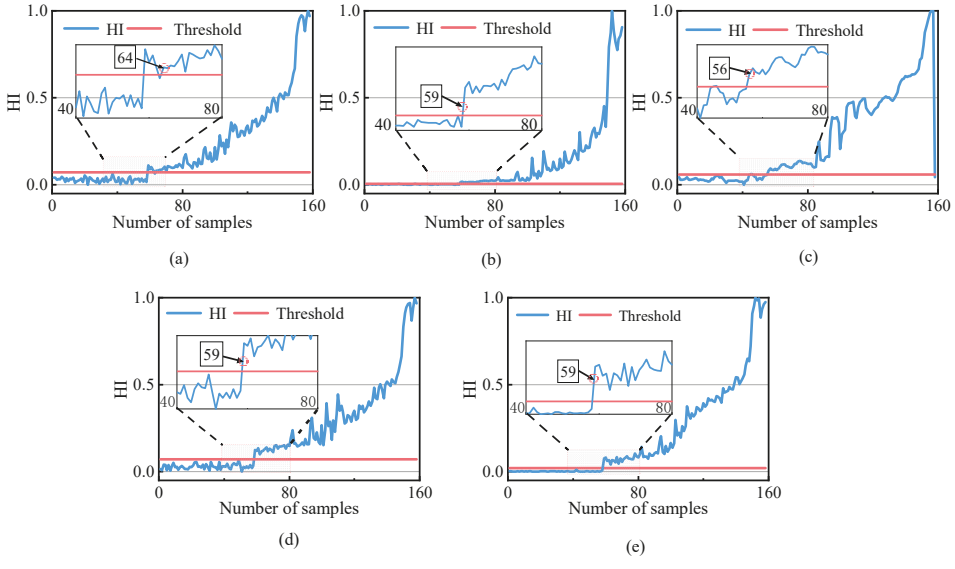


Figure 8. Comparison of the health indicators of different methods. (**a**) PDA results of PCA; (**b**) PDA results of RMS; (**c**) PDA results of CHMM; (**d**) PDA results of SVDD; (**e**) PDA results of the proposed methodology.

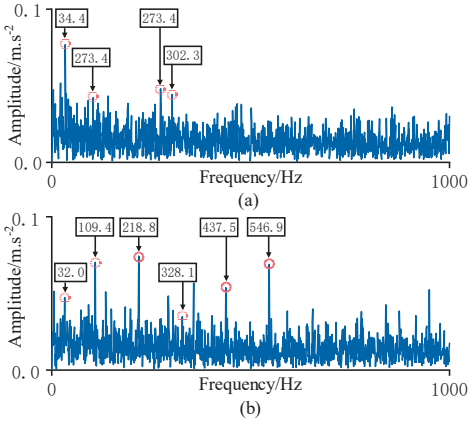


Figure 9. Results of envelope spectrum analysis. (**a**) Envelope spectrum of sample No. 58; (**b**) Envelope spectrum of sample No. 59.

In order to quantitatively evaluate the pros and cons of HIs constructed using different methods, multiple evaluation indicators are used to evaluate the results in this study. In addition to the monotonicity index (Mon), the robustness index (Rob), and the correlation index (Cor) constructed according to Equations (1)–(3), this paper also introduces the separability index (Sep) to quantitatively evaluate the HIs. The Sep was used to measure the ability of the assessment results to discriminate the degradation stage and the ability to warn of the early failure of rolling bearings [1,30]. TOPSIS was used to combine the evaluation indicators to comprehensively measure the results of the evaluation.

The evaluation index and FPT determined using each method are displayed in Table 2. From Table 2, it can be seen that for the HI constructed using the proposed method on this

dataset, only the Cor was slightly lower than the score of the CHMM method. However, the Sep, Mon, and Rob indexes had the highest score, and our proposed model had the best comprehensive score.

Table 2. Evaluation results of the different health indicators.

Index Method	FPT	Sep	Mon	Cor	Rob	TOPSIS Score
PCA	64	0.703	0.981	0.828	0.989	0.621
RMS	59	0.370	0.974	0.801	0.773	0.325
CHMM	56	0.563	0.912	0.953	0.898	0.532
SVDD	59	0.523	0.967	0.824	0.881	0.477
Proposed method	59	0.900	0.986	0.927	0.922	0.914

The local magnification of each evaluation method showed that the HI of the proposed method was the smoothest in the healthy phase, while the HI of other methods fluctuates greatly. In addition, the HI of the proposed method had the highest robustness and comprehensive score. Moreover, the envelope spectrum analysis shows that the proposed method accurately determines the FPT, which provides an adequate warning for equipment maintenance. Therefore, it can be seen that the proposed method is sensitive to the performance degradation of rolling bearings in the whole life cycle and can better tolerate outliers and false fluctuations.

5.2. Case 2: IEEE PHM2012 Data Challenge Dataset

5.2.1. Experimental Description of Case 2

The IEEE PHM2012 Data Challenge dataset provides the full-life vibration signals of rolling bearings in both the horizontal and vertical directions [32]. Figure 10 shows the experimental system. The sampling frequency of the vibration signal was 25.6 kHz, the sampling interval was 10 s, and the sampling time was 0.1 s.

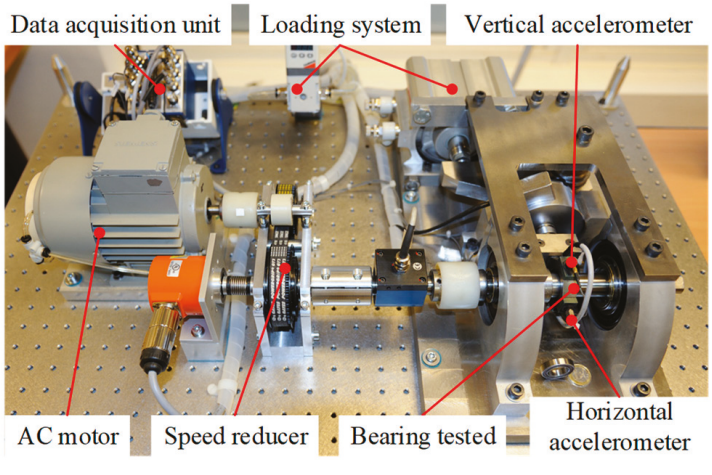


Figure 10. PRONOSTIA platform was used for the PHM2012 datasets.

Verification was carried out using the "Bearing 1.1" subset of the data, which contained 2803 samples. Figure 11 shows the waveform diagram across the time domain for the life cycle of the bearing's data in this test. Then, feature selection and the degradation assessment were carried out using the proposed methods.

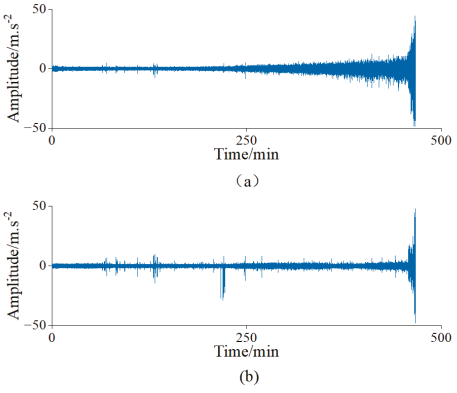


Figure 11. Bidirectional acceleration waveforms of rolling bearings in the time domainof Case 2. (**a**) Horizontal vibration signals; (**b**) Vertical vibration signals.

5.2.2. Experimental Result of Case 2

The results obtained with the methods used for comparison and the proposed methods are shown in Figure 12. In Figure 12a–e, the FPT determined using the proposed method is 190 min, which is earlier than that of the other methods. In Figure 12e, in the first 190 min of operation, the HI value of most samples was below the warning threshold, indicating that the bearing was in a healthy state. After 190 min, the HI value exceeded the threshold value and increased steadily, indicating that the performance of the rolling bearings had begun to deteriorate. Since this dataset does not provide a description of the form of failure, envelope spectrum analysis was not conducted on the samples at the FPT nodes. Local magnification of each evaluation method showed that the HI of the proposed method was the most stable in the healthy phase, and the overall degradation trend was more obvious in the unhealthy state.

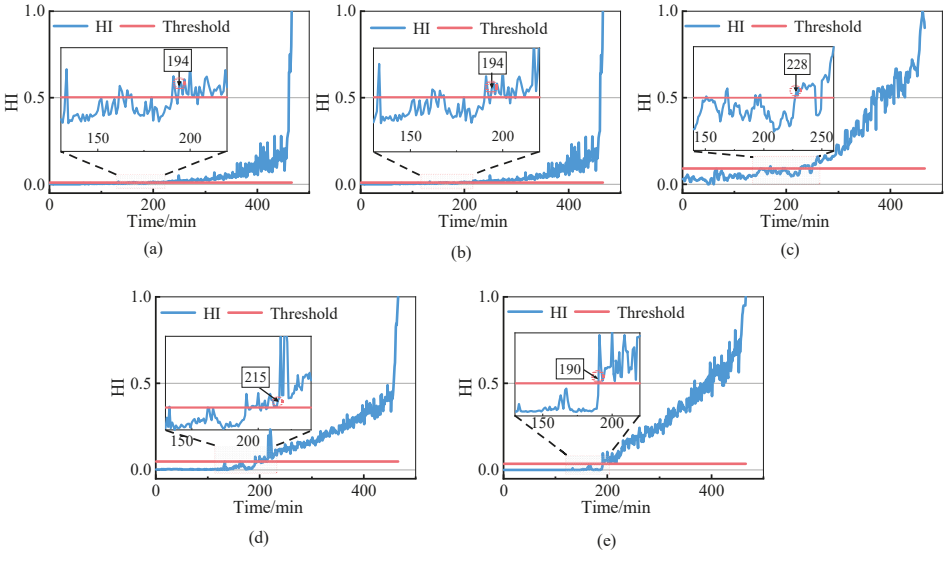


Figure 12. Health indicators of different methods. (**a**) PDA results of PCA; (**b**) PDA results of RMS; (**c**) PDA results of CHMM; (**d**) PDA results of SVDD; (**e**) PDA results of the proposed methodology.

The indexes used for evaluating the health indicators of the previous methods for this subset of data are shown in Table 3. The comparison shows that the health indicators constructed using the proposed method in this study were optimal for all indicators and had the best overall score.

Table 3. Evaluation of the different health indicators.

Index Method	FPT	Sep	Mon	Cor	Rob	TOPSIS Score
PCA	194	0.406	0.965	0.608	0.830	0.097
RMS	194	0.318	0.962	0.608	0.890	0.201
CHMM	228	0.671	0.981	0.955	0.951	0.821
SVDD	215	0.518	0.969	0.866	0.958	0.567
Proposed method	190	0.812	0.987	0.959	0.962	0.998

Furthermore, the local magnification of each evaluation method shows that the HIs of this study's method were the smoothest in the healthy phase, and the HIs of this study's method had the highest robustness score, which shows that this method could better overcome the outliers. In addition, the HIs constructed using the proposed method had the highest comprehensive score, and it determines the FPT earlier, which shows that the proposed method could appropriately reflect the degradation of rolling bearings across their entire life cycle.

5.3. Case 3: Bearing Data from a Home-Made Test Bench

5.3.1. Experimental Description of Case 3

In order to further verify the effectiveness of the method, experimental verification was carried out with a home-made experimental rig for testing accelerated fatigue in rolling bearings. The test bench is displayed in Figure 13. It consisted of an AC motor, a frequency converter, the coupling, the test bearing, the support bearing, a hydraulic loading system, and other components [33]. An SKF-7406 angular contact bearing was used in the experiment. During the experiment, two IMI 603C01 accelerometers were utilized to collect the vertical and horizontal vibration signals of the bearing throughout its life cycle. The sampling frequency of the vibration signal was 25.6 kHz, and the vibration signal was recorded for 1 s every 10 min. The collection was stopped when the maximum amplitude of the collected vibration signal samples exceeded 10 times the maximum amplitude of the initial sample [29].

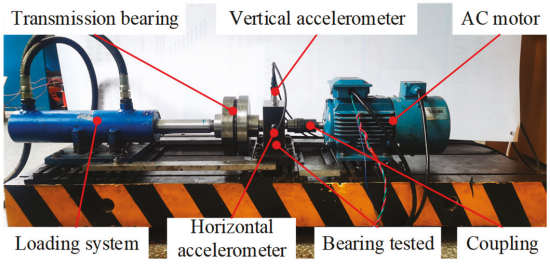


Figure 13. The home-made setup used for the accelerated degradation test.

To maintain the consistency of the experimental conditions, the test bearings were run to failure under a constant load and constant speed. The vibration acceleration data collected for the bearings' entire life cycle from the healthy state to severe degradation are shown in Figure 14. In addition, the proposed method was also used for feature selection and assessing the degradation with the data obtained from the experiments.

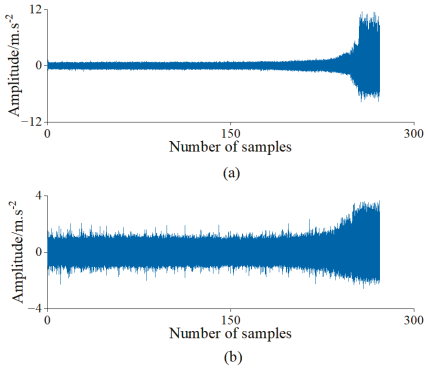


Figure 14. Bidirectional acceleration waveforms of rolling bearings in the time domainof Case 3. (**a**) Horizontal vibration signals; (**b**) Vertical vibration signals.

5.3.2. Experimental Result of Case 3

The results obtained using the proposed method and the methods used for comparison are presented in Figure 15, and the index values and FPT determined using each method are shown in Table 4. As can be observed in Figure 15e, during the first 145 sample periods of operation, the HI values were all below the warning threshold, indicating that the bearing was in a healthy state. After the 145th sample, the HI value continuously exceeded the threshold and gradually increased, indicating that the performance of rolling bearings had begun to deteriorate. Local magnification of each evaluation method showed that the HI of the other methods had false alarm values in the healthy phase. In contrast, the HI of this method had no false alarm values and had minimal fluctuations.

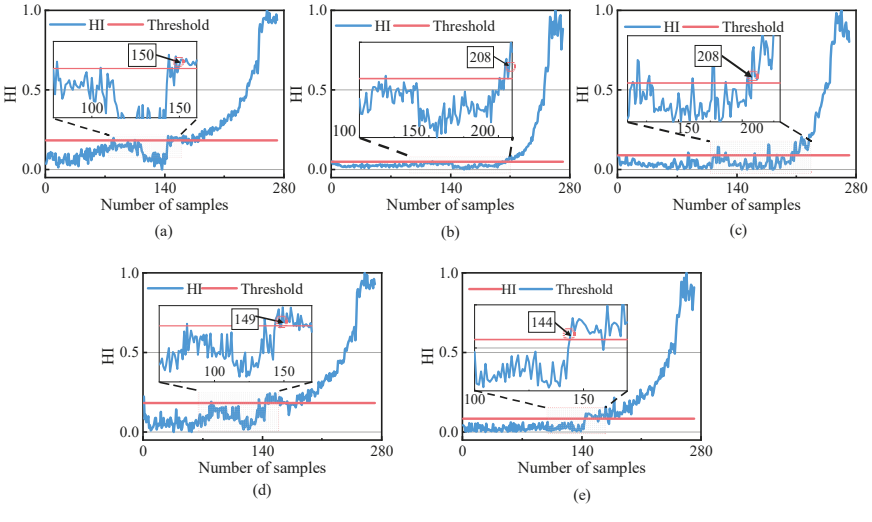


Figure 15. Comparison of HIs of different methods. (**a**) PDA results of PCA; (**b**) PDA results of RMS; (**c**) PDA results of CHMM; (**d**) PDA results of SVDD; (**e**) PDA results of the proposed methodology.

Table 4. Evaluation of the different health indicators.

Index Method	FPT	Sep	Mon	Cor	Rob	TOPSIS Score
PCA	150	0.660	0.964	0.898	0.892	0.626
RMS	208	0.535	0.962	0.859	0.812	0.407
CHMM	208	0.582	0.917	0.933	0.864	0.500
SVDD	149	0.602	0.927	0.863	0.884	0.519
Proposed method	144	0.671	0.965	0.884	0.892	0.835

For these data, the evaluation indicators of the results of the methods mentioned above are listed in Table 4. Table 4 shows that the Cor index of the HI constructed using the proposed method was slightly lower than that of the CHMM method, but the Sep, Mon, and Rob index scores were the best, and our method had the best comprehensive score.

The proposed method had the best comprehensive score, and the FPT was determined earlier than other methods. Meanwhile, local magnification showed that the HI of this method had no false alarm values and had minimal fluctuations, and the HI had the highest robustness score, indicating that the model could better overcome the influence of outliers. Therefore, this showed a good agreement between the results of the degradation assessment and the degree of failure, accurately reflecting the health status of the bearing.

6. Conclusions

Determining the sensitive features set relies heavily on the prior knowledge of experts and degradation models having low-tolerance outliers and false fluctuations, and a method for evaluating the degradation of rolling bearings using adaptive sensitive feature selection and multi-strategy optimized SVDD was proposed. The effectiveness of the method was proved using experiments, leading to the following conclusions:

(1) The TOPSIS–K-medoids method was proposed for adaptive determination of the sensitive feature set. This method determines the adaptive sensitive feature set by using the monotonicity, correlation, and robustness indexes for evaluation, and the process does not need to rely on a priori knowledge to subjectively determine parameters such as the weights and thresholds, which improves the quality of the input data used for the PDA model.

(2) The multi-strategy optimized SVDD strategy trained the model using only the early samples of the healthy phase, adaptively determined the FPT, overcame the interference of outliers and false fluctuations, and better characterized the bearings' degree of failure. The HI showed better consistency with the development trend of faults.

(3) After verification with the XJTU-SY bearing data, the IEEE PHM2012 Data Challenge dataset for bearings, and data obtained with a self-made test bench of accelerated fatigue in rolling bearings, the multi-strategy optimized SVDD model proposed in this paper demonstrated better performance compared to multiple mainstream methods according to a comparison of multiple evaluation indexes, such as monotonicity, correlation, robustness, and separability.

In summary, a rolling bearing performance degradation assessment method with the combination of adaptive sensitive feature selection and multi-strategy optimized SVDD was proposed in this paper. The proposed feature selection method determines the adaptive sensitive feature set with multiple feature evaluation indexes instead of prior knowledge; the multi-strategy optimized SVDD only uses the early samples in the healthy stage to train the model and adaptively determines the FPT while better overcoming the interference of outliers and false fluctuations. The proposed model could accurately reflect the degradation status of rolling bearings verified using experiments, which has a positive effect on the early detection of potential failure of rolling bearings and their maintenance.

Author Contributions: Conceptualization, Z.F. and Z.W.; Data curation, Z.F., Z.W. and J.L.; Formal analysis, Z.F. and Z.W.; Funding acquisition, Z.W.; Investigation, Z.F. and Z.W.; Methodology, Z.F.;

Project administration, Z.W.; Supervision, Z.F. and Z.W.; Validation, Z.W. and X.L.; Visualization, Z.F.; Writing—original draft, Z.F.; Writing—review & editing, Z.F. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 52165065) and the Key Scientific Research Projects of Yunnan Province (Grant No. 202102AC080002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [CrossRef]
2. Schmidt, S.; Heyns, P.S.; Gryllias, K.C. An informative frequency band identification framework for gearbox fault diagnosis under time-varying operating conditions. *Mech. Syst. Signal Process.* **2021**, *158*, 107771. [CrossRef]
3. Yu, H.; Li, H. Pump remaining useful life prediction based on multi-source fusion and monotonicity-constrained particle filtering. *Mech. Syst. Signal Process.* **2022**, *170*, 108851. [CrossRef]
4. Li, Y.; Liang, X.; Lin, J.; Chen, Y.; Liu, J. Train axle bearing fault detection using a feature selection scheme based multi-scale morphological filter. *Mech. Syst. Signal Process.* **2018**, *101*, 435–448. [CrossRef]
5. Wang, Z.; Huang, H.; Wang, Y. Fault diagnosis of planetary gearbox using multi-criteria feature selection and heterogeneous ensemble learning classification. *Measurement* **2021**, *173*, 108654. [CrossRef]
6. Yi, J.; Fang, Z.; Yang, G.; He, S.; Gao, S. New feature analysis-based elastic net algorithm with clustering objective function. *Knowl. -Based Syst.* **2022**, *258*, 110004. [CrossRef]
7. Long, J.; Chen, Y.; Yang, Z.; Huang, Y.; Li, C. A novel self-training semi-supervised deep learning approach for machinery fault diagnosis. *Int. J. Prod. Res.* **2022**, 1–14. [CrossRef]
8. Shao, H.; Xia, M.; Han, G.; Zhang, Y.; Wan, J. Intelligent Fault Diagnosis of Rotor-Bearing System Under Varying Working Conditions with Modified Transfer Convolutional Neural Network and Thermal Images. *IEEE Trans. Ind. Inform.* **2021**, *17*, 3488–3496. [CrossRef]
9. Zhang, Y.; Li, X.; Gao, L.; Chen, W.; Li, P. Ensemble deep contractive auto-encoders for intelligent fault diagnosis of machines under noisy environment. *Knowl. -Based Syst.* **2020**, *196*, 105764. [CrossRef]
10. Xia, J.; Li, Z.; Gao, X.; Guo, Y.; Zhang, X. Real-Time Sensor Fault Identification and Remediation for Single-Phase Grid-Connected Converters Using Hybrid Observers with Unknown Input Adaptation. *IEEE Trans. Ind. Electron.* **2023**, *70*, 2407–2418. [CrossRef]
11. Rai, A.; Upadhyay, S.H. An integrated approach to bearing prognostics based on EEMD-multi feature extraction, Gaussian mixture models and Jensen-Rényi Divergence. *Appl. Soft Comput.* **2018**, *71*, 36–50. [CrossRef]
12. Pan, Y.; Chen, J.; Li, X. Bearing performance degradation assessment based on lifting wavelet packet decomposition and fuzzy c-means. *Mech. Syst. Signal Process.* **2010**, *24*, 559–566. [CrossRef]
13. Wang, H.; Ni, G.; Chen, J.; Qu, J. Research on rolling bearing state health monitoring and life prediction based on PCA and Internet of things with multi-sensor. *Measurement* **2020**, *157*, 107657. [CrossRef]
14. Liu, C.; Gryllias, K. A semi-supervised Support Vector Data Description-based fault detection method for rolling element bearings based on cyclic spectral analysis. *Mech. Syst. Signal Process.* **2020**, *140*, 106682. [CrossRef]
15. Yang, C.; Ma, J.; Wang, X.; Li, X.; Li, Z.; Luo, T. A novel based-performance degradation indicator RUL prediction model and its application in rolling bearing. *ISA Trans.* **2022**, *121*, 349–364. [CrossRef]
16. Song, H.; Kim, M.; Park, D.; Shin, Y.; Lee, J.G. Learning from Noisy Labels with Deep Neural Networks: A Survey. In *IEEE Transactions on Neural Networks and Learning Systems*; IEEE: New York, NY, USA, 2022; pp. 1–19. [CrossRef]
17. Wu, X.; Liu, S.; Bai, Y. The manifold regularized SVDD for noisy label detection. *Inf. Sci.* **2023**, *619*, 235–248. [CrossRef]
18. Wu, J.; Wu, C.; Cao, S.; Or, S.W.; Deng, C.; Shao, X. Degradation Data-Driven Time-To-Failure Prognostics Approach for Rolling Element Bearings in Electrical Machines. *IEEE Trans. Ind. Electron.* **2019**, *66*, 529–539. [CrossRef]
19. Javed, K.; Gouriveau, R.; Zerhouni, N.; Nectoux, P. Enabling Health Monitoring Approach Based on Vibration Data for Accurate Prognostics. *IEEE Trans. Ind. Electron.* **2015**, *62*, 647–656. [CrossRef]
20. Liu, L.; Yu, D. Density Peaks Clustering Algorithm Based on Weighted k-Nearest Neighbors and Geodesic Distance. *IEEE Access* **2020**, *8*, 168282–168296. [CrossRef]
21. Tax, D.M.J.; Duin, R.P.W. Support vector domain description. *Pattern Recognit. Lett.* **1999**, *20*, 1191–1199. [CrossRef]
22. Louhichi, S.; Gzara, M.; Abdallah, H.B. A density based algorithm for discovering clusters with varied density. In Proceedings of the 2014 World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, Tunisia, 17–19 January 2014; pp. 1–6.

23. Guo, W.; Wang, Z.; Hong, S.; Li, D.; Yang, H.; Du, W. Multi-kernel Support Vector Data Description with boundary information. *Eng. Appl. Artif. Intell.* **2021**, *102*, 104254. [CrossRef]
24. Xian, H.; Che, J. Unified whale optimization algorithm based multi-kernel SVR ensemble learning for wind speed forecasting. *Appl. Soft Comput.* **2022**, *130*, 109690. [CrossRef]
25. Okhli, K.; Jabbari Nooghabi, M. On the contaminated exponential distribution: A theoretical Bayesian approach for modeling positive-valued insurance claim data with outliers. *Appl. Math. Comput.* **2021**, *392*, 125712. [CrossRef]
26. Baraldi, P.; Di Maio, F.; Turati, P.; Zio, E. Robust signal reconstruction for condition monitoring of industrial components via a modified Auto Associative Kernel Regression method. *Mech. Syst. Signal Process.* **2015**, *60–61*, 29–44. [CrossRef]
27. Brandsæter, A.; Vanem, E.; Glad, I.K. Efficient on-line anomaly detection for ship systems in operation. *Expert Syst. Appl.* **2019**, *121*, 418–437. [CrossRef]
28. Zhang, J.; Zhang, Q.; Qin, X.; Sun, Y. A two-stage fault diagnosis methodology for rotating machinery combining optimized support vector data description and optimized support vector machine. *Measurement* **2022**, *200*, 111651. [CrossRef]
29. Wang, P.; Long, Z.; Wang, G. A hybrid prognostics approach for estimating remaining useful life of wind turbine bearings. *Energy Rep.* **2020**, *6*, 173–182. [CrossRef]
30. Song, Y.; Liu, D.; Hou, Y.; Yu, J.; Peng, Y. Satellite lithium-ion battery remaining useful life estimation with an iterative updated RVM fused with the KF algorithm. *Chin. J. Aeronaut.* **2017**, *31*, 31–40. [CrossRef]
31. Li, N.; Lei, Y.; Lin, J.; Ding, S.X. An Improved Exponential Model for Predicting Remaining Useful Life of Rolling Element Bearings. *IEEE Trans. Ind. Electron.* **2015**, *62*, 7762–7773. [CrossRef]
32. Nectoux, P.; Gouriveau, R.; Medjaher, K.; Ramasso, E.; Chebel-Morello, B.; Zerhouni, N.; Varnier, C. *PRONOSTIA: An Experimental Platform for Bearings Accelerated Degradation Tests*; IEEE: New York, NY, USA, 2012; pp. 1–8.
33. Wang, Z.; Wu, X.; Liu, X.; Cao, Y.; Xie, J. Research on feature extraction algorithm of rolling bearing fatigue evolution stage based on acoustic emission. *Mech. Syst. Signal Process.* **2018**, *113*, 271–284. [CrossRef]

sensors

MDPI

Article

Fault Monitoring Based on the VLSW-MADF Test and DLPPCA for Multimodal Processes

Shu Wang *, Yicheng Wang, Jiarong Tong and Yuqing Chang

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
* Correspondence: ws_ctrl_engr@163.com

Abstract: Actual industrial processes often exhibit multimodal characteristics, and their data exhibit complex features, such as being dynamic, nonlinear, multimodal, and strongly coupled. Although many modeling approaches for process fault monitoring have been proposed in academia, due to the complexity of industrial data, challenges remain. Based on the concept of multimodal modeling, this paper proposes a multimodal process monitoring method based on the variable-length sliding window-mean augmented Dickey–Fuller (VLSW-MADF) test and dynamic locality-preserving principal component analysis (DLPPCA). In the offline stage, considering the fluctuation characteristics of data, the trend variables of data are extracted and input into VLSW-MADF for modal identification, and different modalities are modeled separately using DLPPCA. In the online monitoring phase, the previous moment's historical modal information is fully utilized, and modal identification is performed only when necessary to reduce computational cost. Finally, the proposed method is validated to be accurate and effective for modal identification, modeling, and online monitoring of multimodal processes in TE simulation and actual plant data. The proposed method improves the fault detection rate of multimodal process fault monitoring by about 14% compared to the classical DPCA method.

Keywords: multimode process; mode identification; process monitoring; statistical modeling

Citation: Wang, S.; Wang, Y.; Tong, J.; Chang, Y. Fault Monitoring Based on the VLSW-MADF Test and DLPPCA for Multimodal Processes. *Sensors* **2023**, *23*, 987. https://doi.org/10.3390/s23020987

Academic Editors: Dong Wang, Jongmyon Kim, Shilong Sun and Changqing Shen

Received: 17 November 2022
Revised: 11 January 2023
Accepted: 12 January 2023
Published: 14 January 2023

1. Introduction

The requirements for safety in industrial production have increased in rigor with the rapid expansion of modern industry. Accurate fault monitoring, diagnosis, and treatment of industrial processes contribute to the normal operation of industrial production and the prevention of further accident spreading. However, the modern industrial structure is complex, with numerous subsystems and variables that are frequently dynamic, non-linear, and highly correlated. Furthermore, due to changes in actual production requirements, industrial processes frequently contain multiple stable and transitional modes. These characteristics of complex industrial processes make monitoring and diagnosing faults more difficult.

In recent decades, data-driven methods [1] have been widely used in the field of industrial technology. This is because a data-driven method does not need to establish an accurate mathematical model or possess too much prior knowledge. Such methods mainly depend on a large amount of process data to analyze and monitor [2–4] the operation of the system or equipment under study. Currently, the mainstream research methods of fault monitoring and diagnosis technology primarily include data-driven multivariate statistics and novel machine learning or artificial intelligence methods. Aldrich and Auret provided a comprehensive review of unsupervised machine learning-based process monitoring methods [5]. Fan trained the autoencoder using offline normal data by building the structure of a neural network and then used it for online fault detection [6].

Classical data-driven approaches include PCA [7], ICA [8], PLS [9], and other methods. Each of these methods has its own strengths and drawbacks. Traditional PCA is more

suitable for production process data that are linear and satisfy Gaussian distribution, while it performs poorly for data that are strongly nonlinear, dynamic, and coupled. Therefore, many scholars have made improvements based on PCA and proposed many methods. For example, Lee proposed KPCA to handle nonlinear data for industrial process data presenting strong nonlinearity [10]. Ku considered the serial correlation of process data and proposed DPCA, which performs better for highly dynamic data [11]. Bakshi proposed MSPCA to solve the multiscale problem [12]. Harrou combined the multivariate exponentially weighted moving average (MEWMA) monitoring scheme with PCA modeling to improve anomaly detection performance [13], and so on.

Considering the importance of local data features represented by data neighborhood information, manifold learning is proposed to provide a novel perspective to preserve the local features of data [14]. According to this method, data are formed by mapping a low-dimensional manifold onto a high-dimensional space. As a result, the low-dimensional data can uniquely represent the original data. Extraction of low-dimensional manifolds from high-dimensional data is accomplished by first establishing a local reduced-dimensional mapping relationship and then attempting to generalize the local mapping relationship to the global. Isomap [15], Laplacian eigenmaps (LE) [16], and locality-preserving projections (LPP) [17] are popular manifold learning techniques. Currently, manifold learning has been applied to process monitoring in industrial processes. By using LPP in combination with PCA, Yu proposed a principal component analysis method (LGPCA) for local and global applications [18]. Luo further revealed the relationship between LPP and PCA and proposed a novel dimensionality reduction algorithm, GLPP, which aims to preserve the global and local structure of the dataset by solving a biobjective optimization function [19]. Wu used PCA, LPP, and isometric feature mapping (ISOMAP) to fuse features extracted from vibration signals for fault diagnosis [20]. These methods have proven to outperform PCA-based and LPP-based monitoring methods.

However, actual industrial process variables are highly dynamic and have characteristics such as autocorrelation and intercorrelation. Traditional methods are difficult to effectively model highly dynamic data, which may result in false alarms in online monitoring. Moreover, due to the changing input point of the working condition and changes in the underlying raw material, the operating state of the industrial process will change to varying degrees, thus showing several different modes. Most data-driven methods operate in a single stable mode but perform poorly on multimodal process data. To accurately model and monitor multimodal processes, some methods have been proposed and practiced in the past. There are two main ideas: 1—Overall modeling [21–23] entails using the same model to describe different modes. 2—Multimodal modeling [24–26] involves describing the process characteristics of each mode by building local models for different stable modes. The goal of overall modeling is to build models that describe the different structures of all modes, such as global PCA models. However, this kind of method can lead to the deterioration of monitoring accuracy for some modes. False alarms may even occur. Multimodal modeling-based approaches model different modes separately. Modeling individual modes is more accurate than overall modeling [27,28].

The main contributions of this paper are as follows: first, an offline mode identification method based on the variable-length sliding window-mean augmented Dickey–Fuller (VLSW-MADF) test is proposed. The commonly used offline mode identification work is based on the trend of variation of each variable for mode classification. The method innovatively uses the smoothness of the data as the basis for stable and transitional mode identification. Compared with other multimodal classification methods, the proposed method in this paper is more intuitive, and the starting position of transition modes can be determined more accurately. Secondly, this paper improves the traditional data-driven fault monitoring method and proposes a novel fault monitoring method DLPPCA. Many scholars have studied the modeling of transition modes [29]; however, these methods often do not focus on the transition modes themselves. DLPPCA performs well on dynamic transition mode data, can accurately model and monitor both transition and stable modes,

and is more suitable for modeling and monitoring multimodal processes. Finally, this paper proposes a novel and less computationally intensive online modal identification method. The traditional online modal identification method requires traversing all offline models [30]. When the number of modalities in the process is large, the computational effort is too large. The modal identification method proposed in this paper is based on matching value calculation and uses an offline matching matrix with the same sample length as the online data for modal identification, which reduces the computational effort and improves the accuracy at the same time.

The paper is organized as follows: Section 2 describes process monitoring based on DLPPCA. Section 3 describes the offline mode recognition and modeling steps based on the VLSW-MADF test and DLPPCA. Section 4 describes the online mode recognition and monitoring strategies proposed in this paper. Section 5 uses the TE process and actual power plant data to simulate and verify the validity and correctness of the methods presented in this paper. Finally, the conclusion of this article is presented in Section 6. To avoid confusion among the many symbols, we have listed a nomenclature.

2. Process Monitoring Based on DLPPCA

PCA is a widely used data-driven method that performs well on data feature extraction tasks and is often applied for process monitoring in industrial practice [31–35]. However, this method often ignores the local structure underlying data, resulting in the loss of potential information from such structures. Locality-preserving projection (LPP) is a manifold learning method that maintains the local structure of data and can restore a low-dimensional manifold structure from high-dimensional sample data [36–38]. At present, scholars use LPP in combination with PCA [39,40], but the statistical model established by this combination method is static; that is, it assumes that the current process is time-invariant. In real industrial processes, process variables have dynamic characteristics of autocorrelation and cross-correlation. Since static PCA is unable to extract dynamic relationships from the data, autocorrelation and cross-correlation are mixed together, which makes it difficult for traditional PCA to reveal what type of relations among the measured variables. Direct application of traditional fault monitoring methods to dynamic data may lead to misleading results (real-time statistics exceeding real-time thresholds, resulting in false fault alarms). Therefore, we must consider the process data serial correlations to implement an efficient monitoring method.

Therefore, a process monitoring method based on dynamic locality-preserving principal component analysis (DLPPCA) is presented to solve the above problems. DLPPCA first constructs an extended matrix to associate adjacent sample points, and this solves the problem of strong correlation among the sample points in a dynamic process. LPP and PCA are combined to extract the maximum variance information of the manifold structure. This algorithm not only solves the problem of traditional data-driven methods having difficulty modeling due to the strong dynamic natures of industrial processes, but also makes up for the disadvantage of PCA or LPP being used alone by combining LPP with PCA. The steps for DLPPCA are as follows:

First, we assume that the sample set is $X \in \mathbb{R}^{n\times m}$ (m is the number of variables, and n is the number of samples) and that the sample set X has been standardized.

$$X = [x_m(1), x_m(2), \ldots, x_m(n)]^T \in \mathbb{R}^{n\times m} \tag{1}$$

The original sample set X is dynamically expanded into a new matrix X^* by adding the time lag values of the variables using the "time lag shift" method proposed by Ku [11]. The sample set: $X \in \mathbb{R}^{n\times m}$ is expanded to:

$$X^* = [X(t), X(t-1), \cdots X(t-l)] \in \mathbb{R}^{(n-l)\times m(l+1)} \tag{2}$$

where l is the number of lags. It is selected by experience and should not be too large; generally, $l = 1, 2$. Where $X(t)$ is the first column of the dynamic expansion matrix and $x^T(t)$ is the m-dimensional observation vector in the sample set at moment t.

$$X(t) = [x^T(t), x^T(t-1), \cdots, x^T(t-n+l)]^T \tag{3}$$

Next, the low-dimensional manifold structure of the data is extracted. The goal of manifold feature extraction is to find a projection matrix $A = [\alpha_1, \alpha_2, \ldots, \alpha_k]$ $k < m(l+1)$ such that the extracted low-dimensional manifold $F = [f_1, f_2, \ldots, f_n] \in \mathbb{R}^{k\times(n-l)}$ retains a local structure similar to that of X^*. Then, we have the following objective function:

$$\begin{aligned} X\tfrac{1}{2} min \sum_{i,j}^{n} (f_i - f_j)^2 W_{ij} \\ = \tfrac{1}{2} min \sum_{i,j}^{n} (A^T X(i) - A^T X(j))^2 W_{ij} \\ = min \sum_{i}^{n} A^T X(i) D_{ii} X(i)^T A - min \sum_{i,j}^{n} A^T X(i) W_{ij} X(j)^T A \\ = min A^T X^* (D - W) X^{*T} A \end{aligned} \tag{4}$$

where W is a W_{ij} relational matrix

$$W_{ij} = \begin{cases} e^{-\frac{\|X(i)-X(j)\|^2}{t}} & X(i) \in N_k(X(j)) \text{ or } X(j) \in N_k(X(i)) \\ 0 \quad \text{otherwise} \end{cases} \tag{5}$$

and D is an $n \times n$ diagonal matrix. The diagonal elements are $D_{ii} = \sum_{j=1}^{n} W_{ij}$ and can be used to indicate the importance of each sample. To ensure that the objective function is solvable, a restriction $FDF^T = I$ or $A^T X^* D X^{*T} A = I$ needs to be added. We define a Laplacian matrix L_P: $L_P = D - W$. Equation (4) is converted to an optimization problem, as shown below:

$$min A^T X^* L_P X^{*T} A s.t. A^T X^* D X^{*T} A = I \tag{6}$$

Equation (6) is equivalent to solving the generalized eigenvalue problem shown below:

$$X^* L_P X^{*T} \alpha = \lambda X^* D X^{*T} \alpha s.t. A^T X^* D X^{*T} A = I \tag{7}$$

The projection matrix A is composed of the eigenvectors corresponding to the k minimum generalized eigenvalues obtained. Thus, the extracted low-dimensional manifold $F = A^T X^*$ is obtained.

Finally, the principal components of manifold F are extracted by PCA. The covariance matrix $\sum$ of the low-dimensional manifold F is as follows:

$$\sum = cov(F) = \frac{F(F)^T}{n - l - 1} \tag{8}$$

Eigenvalue decomposition is performed on the resulting covariance matrix as follows:

$$\sum p_i = \lambda^* p_i \tag{9}$$

The projection matrix P is obtained by taking the eigenvectors corresponding to the d largest eigenvalues. Finally, the feature data extracted by DLPPCA are obtained:

$$T = P^T F = P^T A^T X^* \tag{10}$$

The feature data $T \in \mathbb{R}^{d\times(n-l)}$ from Equation (10) are called the matching matrix. This matching matrix will be used later for online modal recognition.

The above derivation explains the basic principles of DLPPCA. Statistics and confidence limits also need to be built to monitor an industrial process. This paper is im-

plemented with the T^2 and SPE statistics. Among them, the T^2 statistic represents the fluctuations of model variables, and the SPE statistic measures the goodness of fit of the constructed model. Once the statistics of the online data exceed the corresponding confidence limits calculated from the normal offline data, the current process is considered to have a fault situation.

The T^2 statistic is as follows:

$$T^2 = f^T P \Lambda^{-1} P^T f \tag{11}$$

where $f^{(k\times 1)} \in F$ and Λ is a diagonal matrix composed of the largest d in λ^*.

The T^2 statistic obeys the F distribution, so the confidence limit for T^2 is:

$$T_\alpha{}^2 = \frac{k((n-l)^2-1)}{(n-l)(n-k)} F_\alpha(k, n-l-k) \tag{12}$$

The SPE statistic is as follows:

$$SPE = (f)^T \left(I - PP^T\right) f \tag{13}$$

The confidence limit for the SPE is:

$$SPE_\alpha = \theta_1 [\frac{c_\alpha h_0 \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1{}^2}]^{\frac{1}{h_0}} \tag{14}$$

where

$$\theta_r = \sum_{j=k+1}^{m} \lambda^{*r}_j (r = 1, 2, 3) \tag{15}$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2{}^2} \tag{16}$$

$1 - \alpha$ represents the confidence level, which is 0.99 for this article.

3. Offline Mode Identification Based on the VLSW-MADF Test and Modeling

Multimodal processes contain different stationary and transition modes [41]. The stable mode mentioned in this paper refers to the industrial production process in a smooth working condition for a period of time. A stable mode means most of the time that the process data for that time series fluctuates around a stable central level. The nonstationary mode is the state of the production process when it transitions from one operating condition to another. The process data in this time series tend to have a clear upward or downward trend. Moreover, time series with nonstationary states can often be differentiated to form stationary series.

Notably, the sampling data of a multimodal process are also essentially a time series. Therefore, from the perspective of data stationarity, the stable mode can be considered as the current process data is in a stationary state, while the transition mode can be considered as the current process data is in a nonstationary state. From this point of view, this paper presents a method of pattern recognition based on the VLSW-MADF test. The method uses the stationarity of the process data as the basis for the identification of stable and transitional modes. Compared with other multimodal identification methods, the method presented in this paper starts with the intrinsic characteristics of the given data, which is more intuitive and less difficult to implement.

3.1. ADF Test

The augmented Dickey–Fuller (ADF) test is a stability test method that is widely used in the field of economics [42–44]. However, to the authors' knowledge, the ADF test has not yet been applied in the field of process monitoring or fault diagnosis. This method makes a stationarity judgment by determining whether there is a unit root in the current

data; if there is no unit root, the data are in a stationary state. If there is a unit root, the data are in a nonstationary state. The specific process of the ADF test is as follows:

Assume that we have a time series denoted by $\widetilde{X} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n\times 1}$, ($n$ is the number of samples). The ADF test can be completed by validating the following three models after making a first-order difference equation for $\widetilde{X}$:

Model 1:

$$\Delta x_t = \delta x_{t-1} + \sum_{i=1}^{n} \gamma_i \Delta x_{t-1} + \varepsilon_t \tag{17}$$

Model 2:

$$\Delta x_t = \eta + \delta x_{t-1} + \sum_{i=1}^{n} \gamma_i \Delta x_{t-1} + \varepsilon_t \tag{18}$$

Model 3:

$$\Delta x_t = \eta + \beta t + \delta x_{t-1} + \sum_{i=1}^{n} \gamma_i \Delta x_{t-1} + \varepsilon_t \tag{19}$$

where t is the time index, η is an intercept constant called a drift, β is the coefficient on a time trend, γ_i is a trend term, δ is the coefficient presenting process root, and ε_t is a white noise sequence.

The assumptions presented are as follows:

Hypothesis 0 (H0). $\delta = 0(\textit{There is a unit root, and the data are nonstationary})$.

Hypothesis 1 (H1). $\delta < 0(\textit{No unit root and the data are stable})$.

This test is performed by calculating the t-statistic for each model:

$$t_s = \frac{\hat{\delta} - 1}{\hat{\sigma}_\delta} \tag{20}$$

where $\hat{\delta}$ is the estimated value of δ and $\hat{\sigma}_\delta$ is the standard error.

By querying the ADF threshold table, if the obtained t-statistic is less than three confidence levels (10%, 5%, and 1%), it can be judged that the null hypothesis H_0 is rejected with 90%, 95%, and 99% confidence, respectively. If t_s is greater than or equal to the critical value, the current data are not stationary. If t is less than or equal to the critical value, the current data are stationary.

Since it is not known in the actual test which model the data being tested conform to at this time, the ADF test first checks model 3 (Equation (19)), and then it checks model 2 (Equation (18)) and model 1 (Equation (17)) in turn. If the null hypothesis is rejected, the test stops; otherwise, the test continues. That is, when none of the three models can reject the null hypothesis, the time series tested is considered nonstationary, and if one model rejects the null hypothesis, the time series is considered stationary.

3.2. Mode Identification Based on the VLSW-MADF Test

The traditional ADF test introduced in Section 3.1 can only test the stationarity of a single variable. The production data from actual industrial processes are multivariable. As a result, the ADF test cannot be directly applied to industrial process data. However, in pattern recognition, the work that must be carried out is to recognize a pattern according to the changing trend of each variable. If we can extract a single variable that can represent the fluctuation trend of the multivariable industrial data, we can carry out pattern recognition through a stationarity analysis of the single variable. The single variable that can achieve this effect is called the trend variable of the process. This paper proposes a method of using mean value processing to extract the trend variables of a given process. We find that when the process is in a stable mode, the mean value of each sampling point also remains relatively stable; when the process is in a transitional mode, the mean value of

each sampling point likewise fluctuates. The trend variables of the process extracted by the mean processing method can reflect the change trend of the process data accurately. The method for extracting the trend variable is described in detail in the second half of this section.

However, the MADF test alone cannot realize mode identification for process data. If we use the MADF test directly on a whole dataset with multiple modes without distinguishing between them, we will obtain incorrect test results, which cannot be reflected when the process enters a new mode. Only after the process data are divided can mode identification be realized by the MADF test. The result of mode identification is closely related to the length of the selected partition. Therefore, this paper combines the MADF test with a variable-length sliding window and finally proposes the VLSW-MADF test for modal identification. The approximate framework of the method is as follows: First, a window of length H is used to divide the trend variables of the given data, and then the ADF test is used for rough mode identification. Rough mode identification can be used to roughly distinguish stable modes from transition modes. Then, a window of length L is used to divide the trend variables, and the ADF test is used for detailed mode identification. Detailed mode identification can be used to determine the beginning and end of a transition mode. Finally, mode identification is realized for the multimodal process.

We provide the user with a criterion to select the hyperparameters of the proposed offline method. The parameter L should be chosen to satisfy the length of the minimum transition mode of the current process. According to the experience of modeling multivariate statistical regression methods, the window data should be sampled at least 2–3 times more than the number of variables in order to achieve an effective statistical feature extraction. The parameter H should be chosen to satisfy the length of the minimum stable mode of the current process. Moreover, the window length H should be chosen to be at least two times the window length of L ($H \geq 2L$).

The above procedure is the VLSW-MADF test proposed in this paper. It is worth noting that there is no difference between the final modal recognition results obtained using mean processing before and after sliding window partitioning. However, when the mean value is used to divide the sliding window, the method needs to solve the mean value many times. In the VLSW-MADF test proposed in this paper, mean value processing is used before sliding window partitioning to reduce the number of required calculation steps. That is, a sliding window partition and an ADF test are directly carried out on the trend variables.

More detailed steps are as follows. It is assumed that we have multimodal process data $X = [x_m(1), x_m(2), \ldots, x_m(n)]^T \in \mathbb{R}^{n \times m}$ (n is the number of samples, and m is the number of variables). The mean value of the sample point data $x_m(n) = [\alpha_{n1}, \alpha_{n2,\cdots}, \alpha_{nm}]^T \in \mathbb{R}^{m \times 1}$ is calculated to obtain:

$$\overline{x}_m(n) = \frac{\sum_{i=1}^{m} \alpha_{ni}}{m} \tag{21}$$

By summarizing the results of Equation (21), the trend variable of the process is finally obtained as follows:

$$\overline{X} = [\overline{x}_m(1), \overline{x}_m(2), \ldots, \overline{x}_m(n)] \in \mathbb{R}^{1 \times n} \tag{22}$$

Next, rough mode identification is performed: $\overline{X}$ is segmented along the sampling direction using a sliding window H. In this paper, we choose the window length based on the aforementioned criterion and the characteristics of the actual process. In the two numerical simulation cases of this paper, the window length of L is determined to be 50 and the window length of H is determined to be 100. After cutting, we obtain a series of windows: $\overline{X}_1 = [\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_H] \in \mathbb{R}^{1 \times H}$, $\overline{X}_2 = [\overline{x}_{H+1}, \overline{x}_{H+2}, \ldots, \overline{x}_{2H}] \in \mathbb{R}^{1 \times H} \ldots$. The ADF test in Section 3.1 is used to test the stationarity of the data in each window. Finally, a stationarity matrix is obtained as follows:

$$H^* = [h_1, h_2, \ldots h_H] \tag{23}$$

where $h = \begin{cases} 0 \text{ Data are in a nonstationary state} \\ 1 \text{ Data are in a stable state} \end{cases}$.

The nonstationarity window data correspond to processes in a transition mode, while the stationarity window data correspond to processes in a steady mode. The resulting stationarity matrix H^* shows the result of rough mode identification for a multimodal process. However, at this point, we can only obtain a rough idea of whether the process corresponding to each segment of data is in a stable mode or a transitional mode. To achieve pattern recognition, it is necessary to further determine the positions of the beginning and end of each mode. Therefore, detailed mode identification is also needed.

Detailed mode identification: Based on the results of rough mode identification, for the previous window entering the nonstationary state to the next window ending the nonstationary state mode, the shorter window L is used for pattern recognition and stability testing. For example, assuming that the continuous $(p - q + 1)$ values from h_p to h_q in the stationary matrix H^* are all zero, the sample dataset that needs to be reidentified and retested is $\overline{X}_{new} = \left[\overline{X}_{p-1}, \overline{X}_p, \ldots, \overline{X}_q, \overline{X}_{q+1}\right] \in \mathbb{R}^{1\times(p-q+3)}$.

In this step, we need to choose a shorter window length L than H based on the aforementioned criterion and the characteristics of the actual process. We choose the window length $L = 50$. Similar to the matrix obtained via the previous steps of rough mode identification, the final stationarity matrix is as follows:

$$L^* = [h_{l1}, h_{l2}, \ldots h_{lL}] \tag{24}$$

The starting point of the transition mode can be judged more accurately by matrix L^* than by H^*. By analogy, the other transition modes in the rough pattern recognition results are determined in the same way, and finally, pattern recognition is realized for the multimodal process. The detailed steps of the VLSW-MADF test are shown in Figure 1.

3.3. Offline Modeling

The main idea of this method is to first identify a mode and then model it separately. In the offline modeling stage, the transition modes and stable modes obtained after pattern recognition should be modeled individually. In the second section, we declared that the process monitoring method used in this paper is DLPPCA. This method performs well on dynamic data and can accurately model transition modes with large variation ranges and strong dynamics. Although the stable mode means most of the time that the process data of that time series fluctuate around a stable central level, we still have to consider the serial correlation of the stable mode process data. Therefore, DLPPCA is equally applicable to offline modeling for both stable and transition modes. Therefore, this paper uses DLPPCA to model transition modes and stable modes separately based on mode identification. Algorithm 1 shows the offline modeling phase algorithm.

Algorithm 1. Offline modeling phase

Step 1: Input multimodal process data $X = [x_m(1), x_m(2), \ldots, x_m(n)]^T$;
Step 2: Calculate trend variables $\overline{X}$ by (21) and (22);
Step 3: Divide $\overline{X}$ through a window of length H, obtaining H^* through ADF test;
Step 4: Further divide the shorter window L into $\overline{X}_{new}$, obtaining L^* through ADF test;
Step 5: Obtain $X_{t1}, X_{t2}, \cdots$ by step 3 and step 4, and model them separately using DLPPCA, saving ${T_{\alpha t}}^2$ and $SPE_{\alpha t}$;
Step 6: Similar to Step 5, obtain $X_{s1}, X_{s2}, \cdots$ by step 3 and step 4, and model them separately using DLPPCA, saving ${T_{\alpha s}}^2$ and $SPE_{\alpha s}$.

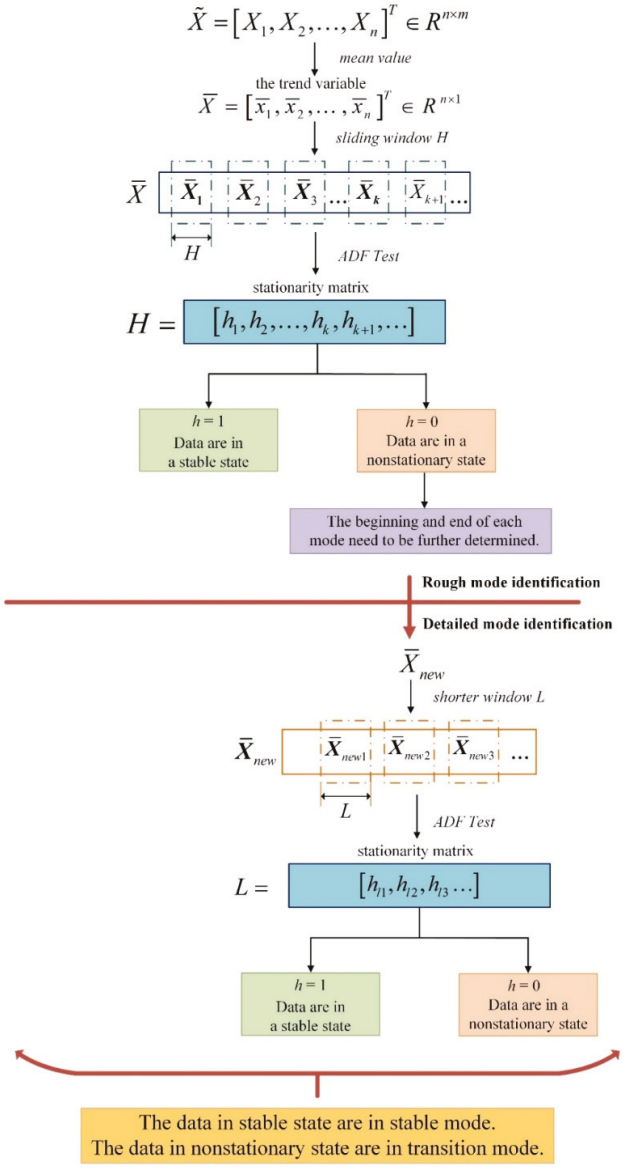


Figure 1. Illustration of the detailed steps of the VLSW-MADF test.

The specific steps are as follows:

Step 1: We acquire multimodal process data $X = [x_m(1), x_m(2), \ldots, x_m(n)]^T \in \mathbb{R}^{n \times m}$.

Step 2: The mean value of the training dataset X is calculated to obtain the trend variables of the process $\overline{X} = [\overline{x}_m(1), \overline{x}_m(2), \ldots, \overline{x}_m(n)] \in \mathbb{R}^{1 \times n}$.

Step 3: The VLSW-MADF test is used to test the stability of $\overline{X}$, determine the starting position of each mode, and complete the pattern recognition task.

Step 4: According to the results of mode division obtained in the previous step, the training data X are divided into several subsegments. The stable mode subsegments are $X_{s1}, X_{s2}, \cdots$, and the transition mode subsegments are $X_{t1}, X_{t2}, \cdots$.

Step 5: DLPPCA is used to model each transition mode subblock. Taking subsegment X_{t1} as an example, according to the content in the first section, the confidence limit ${T_{\alpha t1}}^2$ and $SPE_{\alpha t1}$ can be obtained by using DLPPCA to model X_{t1}, and the final feature extraction result T_{t1} can be obtained. This feature extraction result is also called the matching matrix. This matching matrix and the confidence limit are both saved.

Step 6: Similar to Step 5, DLPPCA is also used to model the stable mode subblocks. Taking subsegment X_{s1} as an example, DLPPCA is used to model and obtain the confidence limit ${T_{\alpha s1}}^2$ and $SPE_{\alpha s1}$. In the subsequent online mode identification step, it is not necessary to use the matching matrices of the stable modes, so only the confidence limit needs to be saved here. A flowchart of the offline modeling is shown in Figure 2.

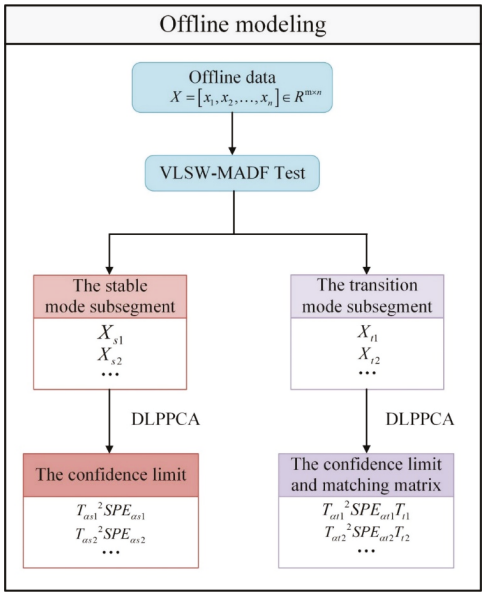


Figure 2. Flowchart of the offline modeling.

It is worth noting that this paper only takes a portion of a process dataset as an example to illustrate the steps of offline mode identification and modeling. However, in practical applications, to model all the modes offline, it is often necessary to identify and model multiple segments of process data. In particular, a transition mode is assumed to contain a stable mode A and stable mode B. The transition process from stable mode A to stable mode B (transition mode AB) and the transition process from stable mode B to stable mode A (transition mode BA) are two different transition modes, and the change trends of their related characteristics are also different, so it is necessary to establish transition models for these two transition modes separately. In other words, if we have stable modes A and B, then A and B satisfy $B \leq A(A-1)$.

4. Online Mode Identification and Monitoring Algorithm

In the last section, offline mode identification and modeling were completed. When conducting online monitoring for multimodal processes, it is also necessary to identify the current online process data. Only by determining which mode the current process belongs to can the appropriate offline model for subsequent online monitoring be selected. If the judgment is wrong, it may result in false alarms. Therefore, when online monitoring is performed, it is necessary to carry out mode identification first and then statistical monitoring.

For online mode identification, researchers have proposed several methods, such as the minimum SPE principle, which involves traversing all models and selecting the

corresponding model with the lowest SPE. Another approach is the probability monitoring method, in which the online samples come from each process with a certain probability, and all offline models are used for joint detection with a certain probability.

However, all offline models need to be considered in the above methods. When there are too many modes in the examined process, the number of calculations is too large. When the process corresponding to the online data first enters the transition mode, the amount of data is small, and the data characteristics are different from those of the whole transition dataset. If the offline model derived from the whole transition dataset is used to match the online data, mode identification errors easily occur.

Therefore, considering the above problems, a new online mode identification method is proposed. The method proposed in this paper does not need to identify all online data but instead discusses them in different situations, and mode identification is performed only in certain situations. The proposed mode identification method is based on the calculation of matching values. An offline matching matrix with the same sample length as that of the online dataset is used for mode identification instead of using the whole transition dataset for matching, thereby improving the accuracy of the method.

It should be noted that since the current sampling time selected for online operation is k, reliable and accurate conclusions cannot be obtained if the online modal recognition process depends only on the results of one sample point at time k. Therefore, online mode identification is performed by combining the recognition results of ω consecutive online sampling data, that is, from the $(k-\omega+1)$ sample to the kth sample. Algorithm 2 shows the online monitoring phase algorithm. Based on the above premises, the detailed steps of the online monitoring method proposed in this paper are as follows (the proposed online mode identification method is used in Step 4).

Algorithm 2. Online monitoring phase.

Step 1: Input online process data X_{online};
Step 2: Determine the mode of the starting phase by minimum SPE;
Step 3: Monitor the current continuous ω data from $(k-\omega+1)$ to k using an offline model corresponding to $(k-\omega)$ time data;
- Situation 1: Below the control limit.
  - The current process mode is the same as the previous one.
- Situation 2: Exceeding the control limit.
  - Situation 2.1: The current process has a fault situation.
  - Situation 2.2: The current process enters a new mode.
    - Situation 2.2.1: The process is in transition mode at the previous moment.
    - Situation 2.2.2: The process is in stable mode at the previous moment.

Step 4: For situation 2.2.2, calculate matching value m_i for online modal recognition.
Step 5: Use the model determined in step 4 to remonitor. If it is below the control limits, the currently selected model matches the actual mode. If it exceeds the control limits, a fault has occurred.

Step 1: Determining the model for the starting stage.

For the initial phase of a process, since there are no data from the sample points at the previous moment to use as references, the corresponding model for the starting phase needs to be determined. Here, the minimum SPE principle is used to determine that for a data segment with a starting length of ω; the online data are monitored in turn using known historically stable modes. The model with the lowest SPE for online samples is selected for monitoring.

Step 2: Trial monitoring of online process data:

When there are $(k-\omega+1)$ consecutive ω data points to k, the offline model from the previous moment is fully utilized for detection. The current continuous ω ($\omega > 10$) data from $(k-\omega+1)$ to k are monitored using an offline model corresponding to $(k-\omega)$ time data.

Step 3: Analyzing the monitoring test results.

There are several possibilities for monitoring the test results obtained in Step 2. If the current process data statistic is below the control limit, it means that the process data ω and $(k-\omega)$ correspond to the same mode of the process. If the current process data statistic exceeds the control limit, the mode changes at the time when the control limit is exceeded. There are two possibilities for change.

1. The current process has a fault situation;
2. The current process enters a new mode. There are also two possibilities for entering a new modal process, from transition mode to stable mode or from stable mode to transition mode.

It should be noted that to avoid false alarms, this paper considers that a process is abnormal only when a continuous number of samples ($\geq \frac{\omega}{2}$) are beyond the control limit and do not depend only on the identification result of a sampling point at time t.

Step 4: Online modal identification.

If the current process data statistics are beyond the confidence limit, it is necessary to judge whether the current process is having a fault or enters a new mode.

First, we assume that the current data enter a new mode.

3. If the process was in the transition mode at the previous moment $(k-\omega)$, the current process enters the stable mode corresponding to the transition mode. This stable mode is selected as the monitoring mode and no mode identification is required;
4. If the previous moment $(k-\omega)$ process is in a stable mode, the current process enters into a transition mode that bridges with this stable mode. Modal identification is required. However, only the transition modes that articulate that stable mode need to be selected for modal identification, not all transition modes need to be selected.

In online mode identification, a method based on matching value calculation is presented in this paper.

Assume that all possible historical transition modes are $X_{t1}, X_{t2}, \cdots$. According to Section 3.3, we can obtain the matching matrix corresponding to the historical transition mode: $T_{t1}, T_{t2}, \cdots$. The online data are modeled by DLPPCA, and the online matching matrix is T_{online}. The historical matching matrix at this time comes from performing feature extraction on the whole transition dataset, and T_{online} is derived only from the current ω data points. The characteristics of transition data are high volatility and a large change range. This means that T_{online}'s data features are different from those of $T_{t1}, T_{t2}, \cdots$. Additionally, direct matching is prone to errors. Therefore, to conduct matching accurately, short processing is performed for $T_{t1}, T_{t2}, \cdots$; that is, the original $T_{ti} \in \mathbb{R}^{d\times(n-l)}$ is truncated along the direction of the sampling point, and only the first ω column vectors are taken. Because the data needed for online mode identification are considered to have just entered the transition mode, it is reasonable to select the first ω column vectors of T_{ti}.

Next, the matching value m_i between the matrices T_{online} and $\widetilde{T_{ti}}$ is calculated in turn. In this paper, the matching value m_i is solved based on Euclidean distance, and the similarity of the two matrices is measured by calculating the sum of the distances between the corresponding column vectors in T_{online} and $\widetilde{T_{ti}}$. The smaller the distance, the smaller m_i is. The specific procedure is as follows:

First, the Euclidean distance vector between T_{online} and $\widetilde{T_{ti}}$ is calculated as follows:

$$D_d = \left[d_1, d_2 \cdots, d_j\right] \tag{25}$$

where

$$d_j = \sqrt{\left(T_{onlinej} - \widetilde{T_{tij}}\right)^T \left(T_{onlinej} - \widetilde{T_{tij}}\right)} \tag{26}$$

Here, $T_{onlinej}$ represents line j of T_{online} and $\widetilde{T_{tij}}$ represents line j of $\widetilde{T_{ti}}$.

The elements of the Euclidean distance matrix D are summed to obtain the matching value m_i, as follows:

$$m_i = d_1 + d_2 + \cdots + d_j \tag{27}$$

The transition mode corresponding to the minimum m_i is selected as the monitoring model.

Step 5: Remonitoring.

The ω consecutive process data points from $(k-\omega+1)$ to k are monitored again using the monitoring model determined in Step 4. The monitoring results are analyzed again. If the current process statistics are below the confidence limit, indicating that the current selection model matches the actual mode, the obtained model can continue to perform process monitoring. If the current data still exceed the confidence limit, a fault has occurred. A flowchart of the online mode identification and monitoring algorithm is shown in Figure 3. To avoid confusion among the many symbols, we have created a nomenclature, as shown in Nomenclature Section.

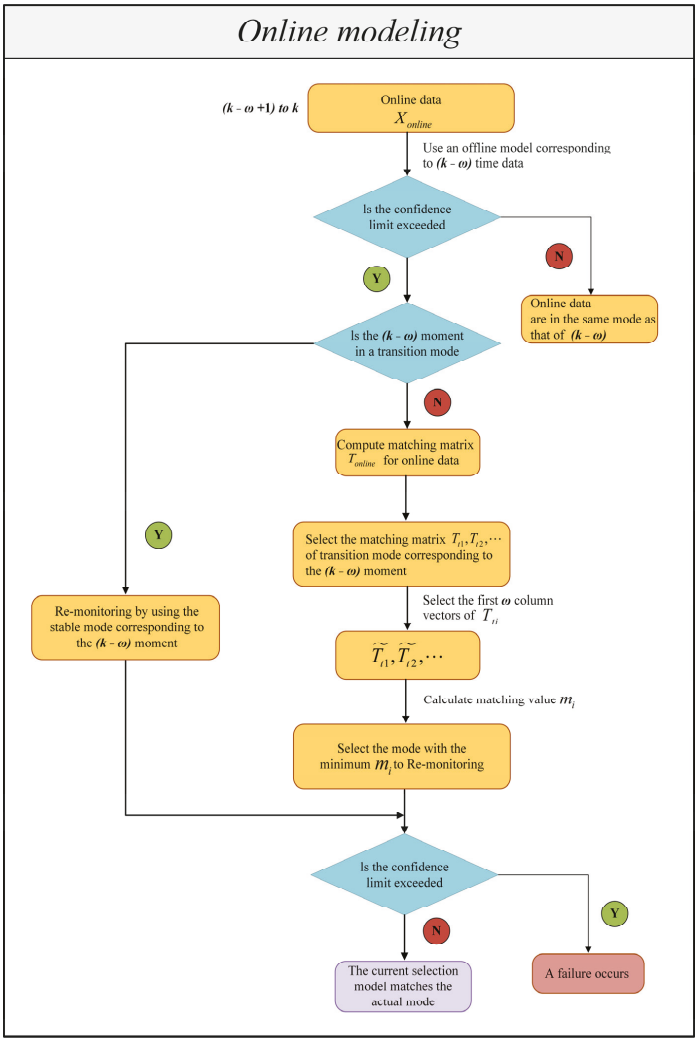


Figure 3. Flowchart of the online modeling.

5. Application and Results

In the previous section, we showed the proposed method in detail. In this section, we will use two numerical simulation cases to verify the effectiveness of our proposed method. First, the first numerical simulation case was carried out based on the TE process. We generated multimodal data based on normal operating conditions by adjusting the operating points of the TE process. Second, the case of the second numerical simulation was carried out based on data from a power plant generating unit. We also simulated a multimodal process data. The validity and feasibility of the methods presented in this paper are verified from the following four perspectives:

1. The presented offline mode identification method based on the VLSW-MADF test is accurate and feasible;
2. The online mode identification method proposed in this paper is accurate and feasible;
3. Transition modes are more accurately modeled and monitored using DLPPCA than with other approaches;
4. Modeling stable modes and transition modes separately can improve the accuracy of online monitoring.

For the proposed method, the fault detection rate (FDR), false alarm rate (FAR), missed alarm rate (MAR), and detection delay (DD) are mainly considered to evaluate the method's performance. These metrics are applied to quantify the method performance in the two subsequent numerical simulation cases. False alarm rate (FAR) measures the probability of false alarms, and a false alarm is an indication of a fault when a fault has not occurred. Fault detection rate (FDR) measures the probability of successful fault detection, and successful fault detection is an indication of a fault when a fault has occurred. Missing alarm rate (MAR) measures the probability of a missed alarm, which is when a fault occurs but is not detected. Detection delay (DD) is the time period between the start of a fault and the time of the detection. It is expected that a larger value for the FDR indicator is better. Smaller values for the remaining three indicators are better. The formulae for calculating the FDR, FAR, and MAR indicators are as follows.

$$\mathrm{FDR} = \frac{\text{number of samples } (I > I_{\mathrm{CL}} \mid \text{fault})}{\text{total samples (fault)}} \times 100\% \tag{28}$$

$$\mathrm{FAR} = \frac{\text{number of samples } (I > I_{\mathrm{CL}} \mid \text{fault} - \text{free})}{\text{total samples (fault} - \text{free)}} \times 100\% \tag{29}$$

$$\mathrm{MAR} = \frac{\text{number of samples } (I \leq I_{\mathrm{CL}} \mid \text{fault})}{\text{total samples (fault)}} \times 100\% \tag{30}$$

where I represents the current data statistic value and I_{CL} represents the control limit. $I = \{T^2, SPE\}$.

5.1. TE Process

A TE process is a simulation based on a real industrial process [45–47]. The operating points of a TE process can be adjusted to meet production requirements when generating multimodal data. This paper describes a 160 h multimodal process; the values of the Production Setpoint, Sep Level Setpoint, and Steam Valve Position are changed at the 50th hour so that the TE process transitions from stable mode A to stable mode B. At 90 h, the values of the production setpoint, sep level setpoint, steam valve position, mole%g setpoint, and yA setpoint are changed again so that the TE process transitions from stable mode B to stable mode C.

Finally, multimodal process data were obtained, with a total of 1600 sample points. This process includes the stable mode A, stable mode B, stable mode C, transition mode AB, and transition mode BC. There are 53 variables in the TE process. Eight process continuous variables are selected to validate the proposed method. These eight variables are shown

in the following Table 1. The change curves of the eight variables of the simulation data under normal working conditions are shown in Figure 4.

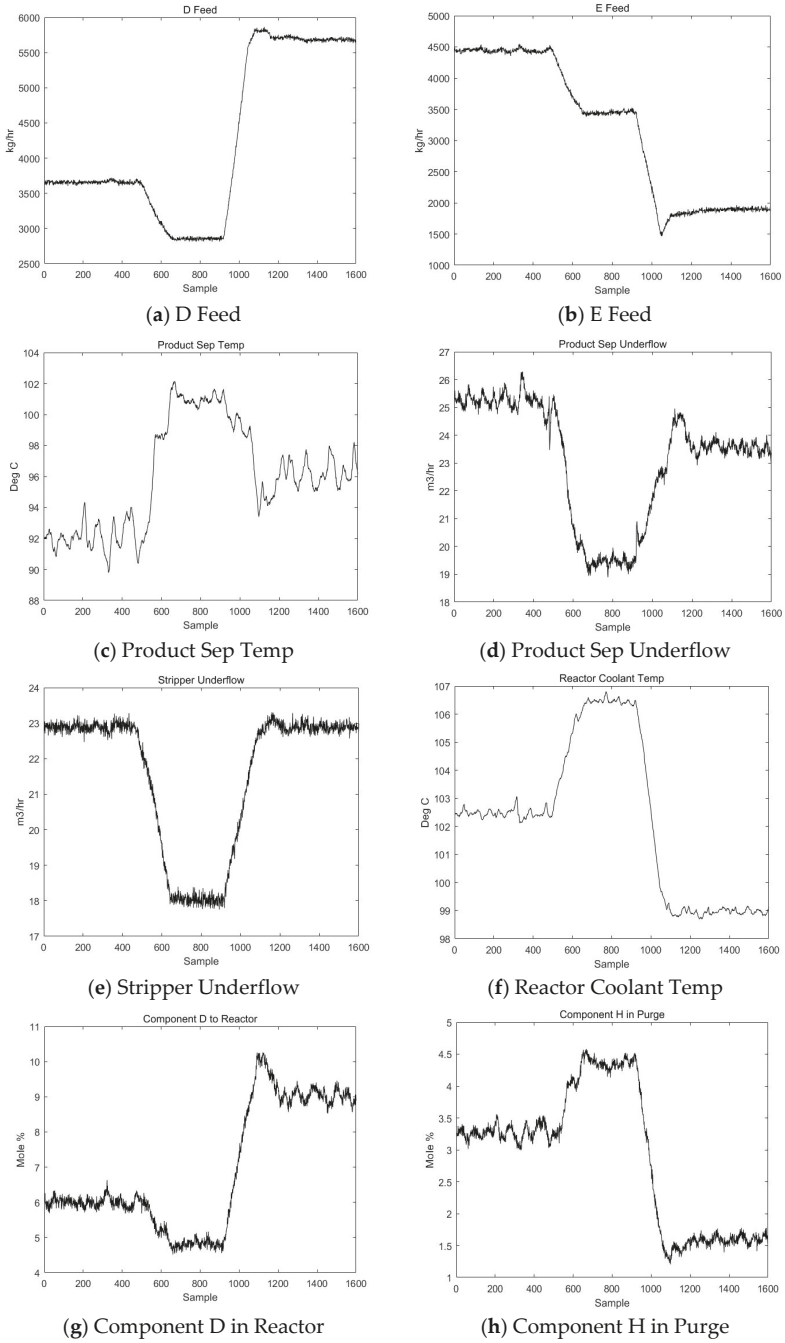


Figure 4. The change curves of the 8 variables of the simulation data.

Table 1. Eight process continuous variables in TE process.

Serial Number	Variable Name
1	D Feed
2	E Feed
3	Product Sep Temp
4	Product Sep Underflow
5	Stripper Underflow
6	Reactor Coolant Temp
7	Component D in Reactor
8	Component H in Purge

The multimodal process dataset consisting of these eight variables is named X_{train}. First, the segment data are identified based on the VLSW-MADF test. The trend variable $\overline{X}$ is derived from Formulas (21) and (22). The change curve for this trend variable is shown in Figure 5. Notably, the trend of $\overline{X}$ coincides with the pattern change trend of the original process; the pattern changes in the 50th and 90th hours and transitions to a new mode each time. This indicates that the variable $\overline{X}$ can represent the trend of multivariable process data.

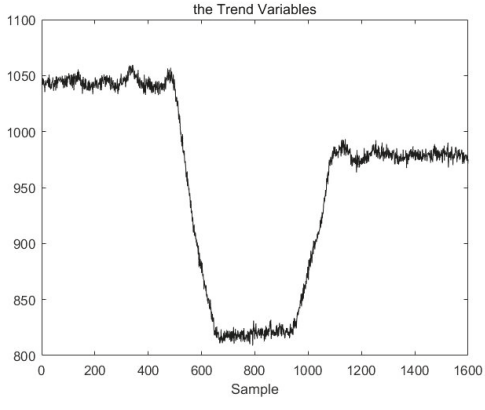


Figure 5. The change curve for this trend variable.

After $\overline{X}$ is obtained, rough mode identification is performed using a window with a length of $H = 100$, resulting in the following stationarity matrix:

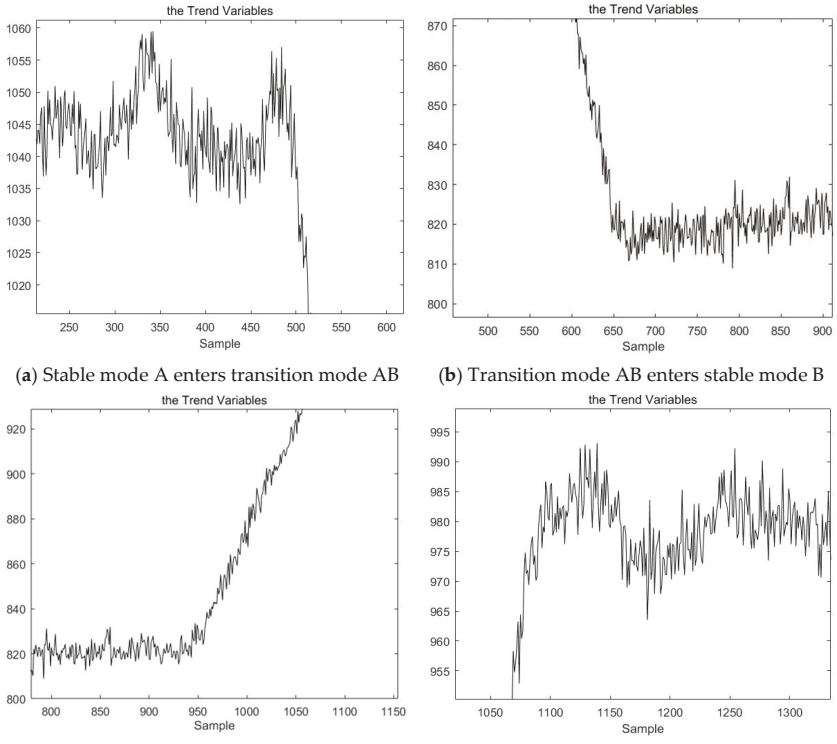
$$H = [1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1]$$

It can be seen from the matrix that the window $\overline{X}_1 - \overline{X}_5$ is in a stable mode. Window $\overline{X}_6 - \overline{X}_7$ is nonstationary and enters a transition mode. Window $\overline{X}_8 - \overline{X}_9$ enters a stable mode. Window $\overline{X}_{10} - \overline{X}_{11}$ enters a transition mode again. The final window $\overline{X}_{12} - \overline{X}_{16}$ remains in a stable mode. In order to highlight the details in the mode transitions, the next step requires more detailed mode identification to determine the exact starting position of the transition modes. Two small stationarity matrices are obtained by dividing the data of windows $\overline{X}_5 - \overline{X}_8$ and $\overline{X}_9 - \overline{X}_{12}$ using a shorter window $L = 50$. The small stationary matrices are shown below:

$$L_1 = [1, 1, 0, 0, 0, 1, 1, 1] L_2 = [1, 1, 1, 0, 0, 0, 1, 1]$$

Finally, results are obtained based on the VLSW-MADF test. The process of data points 1–500 is in stable mode A. The process of data points 500–650 is in transition mode AB. the process of data points 650–950 is in stable mode B. the process of data points 950–1100 is

in transition mode BC. Finally, the process of data points 1100–1600 is in stable mode C. These results are consistent with the actual situation and can be explained with the trend variable $\overline{X}$. Figure 6 is the local magnifications of the trend variable $\overline{X}$ for demonstrating the correctness of the results of mode identification based on the VLSW-MADF test.



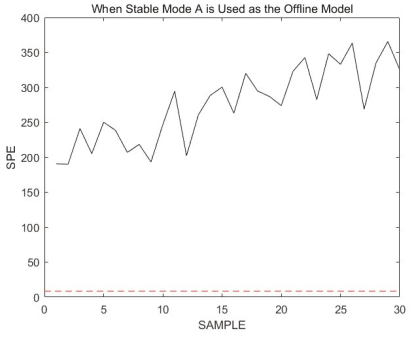
(**a**) Stable mode A enters transition mode AB (**b**) Transition mode AB enters stable mode B

(**c**) Stable mode B enters transition mode BC (**d**) Transition mode BC enters stable mode C

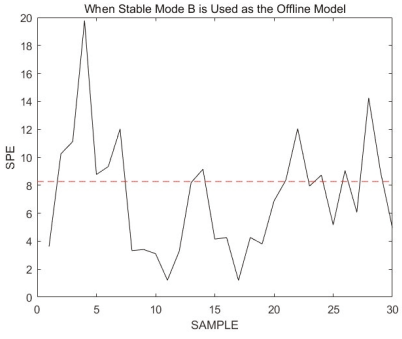
Figure 6. The local magnifications of the trend variable.

Next, based on the modal identification results, stable modes A, B, and C and transition modes AB and BC are modeled using DLPPCA, and the confidence limits and matching matrices of each mode are saved. It is important to note that only a portion of the full dataset is presented here. However, other forms of multimodal data need to be identified and modeled. Finally, the confidence limits and matching matrices of stable modes A, B, and C and transition modes AB, BC, AC, BA, CB, and CA are obtained. In the online monitoring phase, this paper first monitors a section of normal operating process data from stable mode B to stable mode C online. The test data contain a total of 1000 sample points, and the process enters transition mode BC at around the 500th sample point, exits transition mode at around the 700th sample point, and finally enters stable mode C.

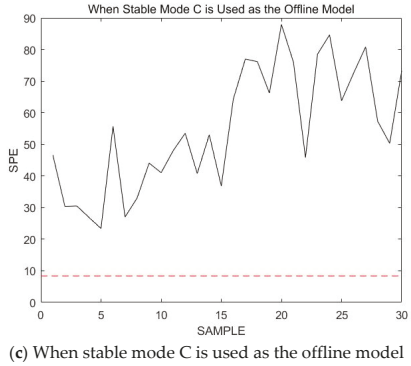
This test dataset X_{test} is used to verify the correctness of the online modal identification method proposed in this paper. As seen in Section 4, there are no previous sample data points available for reference at the beginning of the process. Therefore, the online data are monitored using known stable modes A, B, and C as offline models; that is, 30 consecutive data points from the first sample are monitored online. The obtained results in terms of the SPE statistics are shown in Figure 7.



(**a**) When stable mode A is used as the offline model



(**b**) When stable mode B is used as the offline model



(**c**) When stable mode C is used as the offline model

Figure 7. The obtained results in terms of the SPE statistics.

The red dashed line in the figure represents the confidence limit. Notably, the SPE statistic is the smallest when using stable mode B to detect online process data. From this observation, it is determined that the production process is in stable mode B. This conclusion is consistent with the actual situation. The online process data are then continuously monitored using stable mode B. In subsequent monitoring steps, most of the online data statistics are below the confidence limit, and these online data statistics only occasionally exceed the confidence limit. However, it has been declared previously that a failure or a new mode is only considered if several consecutive sampling points exceed the confidence limit.

When the 521st sampling point to the 550th sampling point is monitored, these 30 consecutive sampling points are beyond the confidence limit, as shown in Figure 8.

At this point, we consider the current process to have transitioned to a new operating mode or to have a fault. The steps in Section 4 are now followed. First, assume that the current process transitions to a new operation mode. Since the process at the previous time is in stable mode B, the current process of this section must be in one of the transition modes connected to B. Therefore, transition mode BA, transition mode BC, and the current process must be selected to match the current mode. According to Step 4 in Section 4, the matching value between the online data and transition mode BA is $m_{BA} = 21.98$, and the matching value for transition mode BC is $m_{BC} = 17.41$. According to these matching values, it is determined that the current process enter transition mode BC. Transition mode BC is used as the offline model to remonitor the current process data online, and the results are shown in Figure 9.

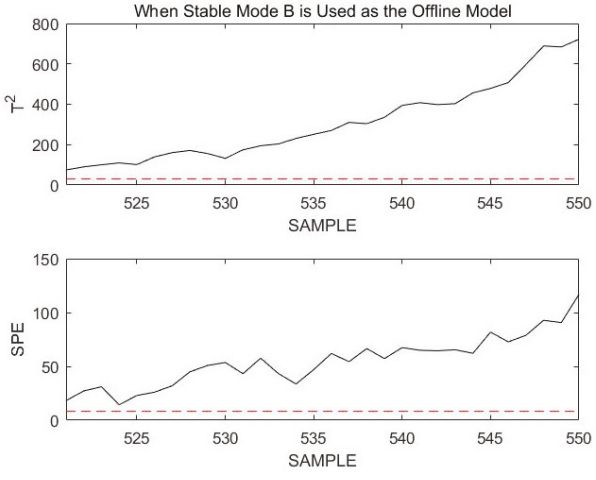


Figure 8. Monitoring 521st to 550th sample points using stable mode B as an offline model.

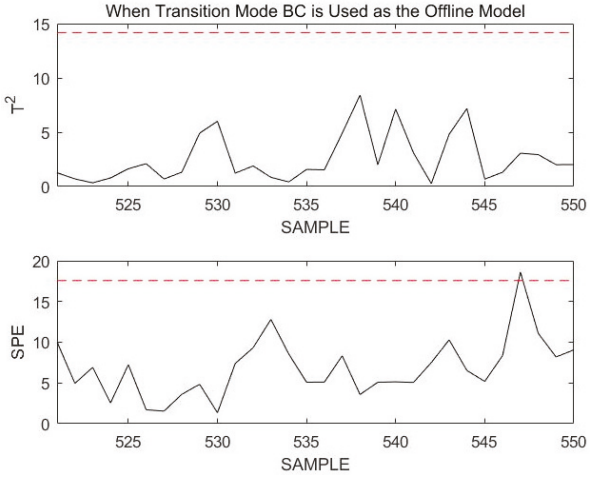


Figure 9. Monitoring 521st to 550th sample points using transition mode BC as an offline model.

At this time, the online monitoring sample statistics are below the confidence limit. It is proven that the current process is in transition mode BC, which is consistent with the actual situation. The above simulation verifies the feasibility and correctness of the online mode identification method proposed in this paper that performs online monitoring and online modal identification on process data obtained under normal working conditions.

The next step is to use a data segment in stable mode A where a fault has occurred online test data. The fault occurs in the variable D Feed; introducing a fault signal at the 51st sample point linearly increases the variable D Feed to simulate a progressively increasing fault. The increased value is maintained between the 80th and 130th sample points; starting at the 131st sample point, the value falls back to its normal level, and at the 150th sample point, the fault disappears. The change curve of D Feed is shown in Figure 10.

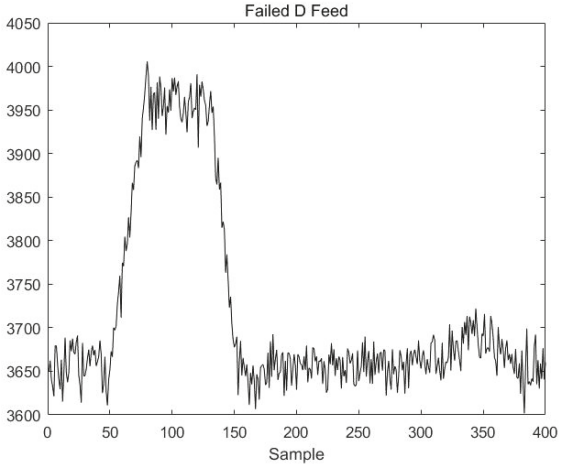


Figure 10. The change curve of D Feed.

The simulation here omits steps such as determining the initial mode and begins directly at the 55th sample point. Since the historical mode of the data at the previous moment is known to be stable mode A, stable mode A continues to be used for the online monitoring of 30 consecutive data points starting at the 55th sample point, as shown in Figure 11.

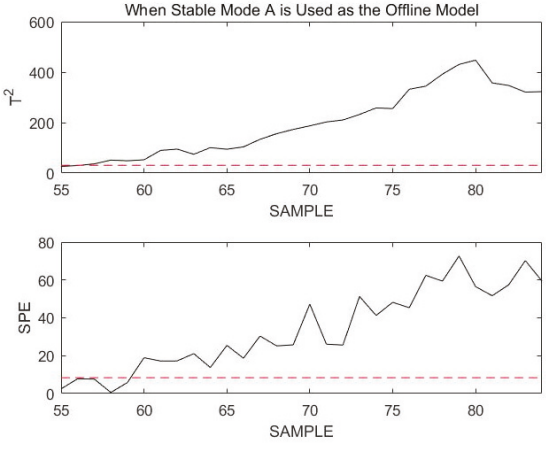


Figure 11. Monitoring 55th to 84th sample points using stable mode A as an offline model.

As seen in the figure above, the T^2 and SPE statistics for the current data almost all exceed the confidence limit. As a result, the current process has faulted or entered a transitional mode. Assuming that the current process enters a transition mode, since the previous moment was stable mode A, the current process can only be in a transition mode joined to A: transition mode AB or transition mode AC. The procedure of Step 4 in Section 4 is continued to obtain the match between the online data and transition mode AB; $m_{AB} = 41.05$, and the matching value for transition mode AC is $m_{AC} = 28.99$. From the matched values, transition mode AC is more likely to be the transition mode in which the online process is located. The online data are remonitored using transition mode AC, as shown in Figure 12.

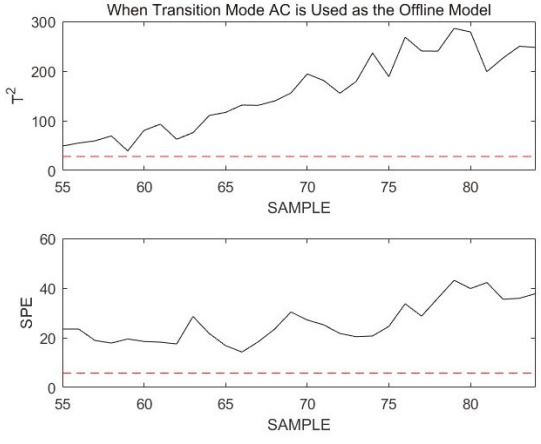


Figure 12. Monitoring 55th to 84th sample points using transition mode AC as an offline model.

Obviously, all data statistics exceed the confidence limit. This indicates that the current process does not enter transition mode AC, but is faulted. Therefore, the current process data will continue to be monitored using stable mode A. To better illustrate the effectiveness of using stable mode A for the failure monitoring of online data, Figure 13 shows the results of online monitoring at sample points 51 to 200.

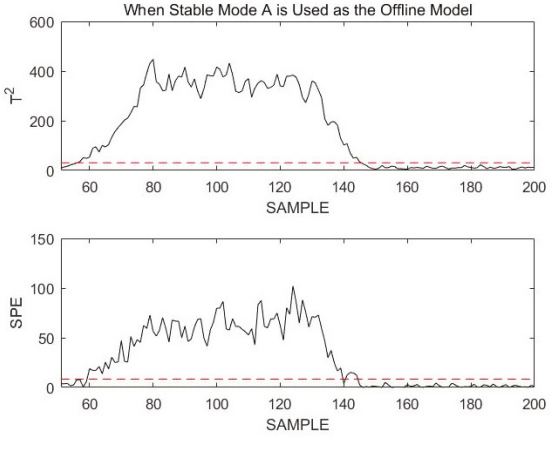


Figure 13. Monitoring 51st to 200th sample points using stable mode A as an offline model.

Figure 13 in the article shows the results of online monitoring based on the modal identification monitoring model. The fault can be clearly detected when T^2 and SPE exceed the control limits at 56 and 60 sampling times, respectively. The fault is introduced from the 51st sampling time. The monitoring statistics T^2 and SPE have a detection delay of 5 sampling times and 9 sampling times, respectively. In addition, the estimated fault end time differs from the real situation by only 4 sampling times, which indicates that the monitoring model DLPPCA based on modal identification can accurately locate the fault interval and has accurate monitoring results. In addition, we calculated FDR, FAR, MAR, and detection delay for the T^2 statistic and SPE statistic, as shown in Table 2.

Table 2. TE simulation case online monitoring results.

Statistics	FDR	FAR	MAR	Detection Delay
T^2	91%	0%	9%	5
SPE	87%	0%	13%	9

The above simulation of the TE process data proves the following:

1. The correctness and feasibility of offline mode identification based on the VLSW-MADF test. The VLSW-MADF method proposed in this paper can accurately and quickly identify the mode of multimodal process.
2. The correctness and feasibility of the online mode identification method proposed in this paper. The online mode identification method proposed in this paper does not require all the online data to be modally identified and makes full use of the data from the previous moment for a case-by-case discussion. When a fault occurs or enters a transition mode, this method can accurately identify.

In Section 5.2, the validation simulation of the actual data from a power plant motor set will be continued to demonstrate the superiority of the DLPPCA method and the necessity of modeling stable modes and transition modes separately.

5.2. Power Plant Data Simulation

In the simulation experiments in this section, relevant data from a 2 × 660 MW power plant are used. A steam feedwater pump system is selected as an example for simulation purposes. The schematic diagram of the thermal power unit is shown in Figure 14. The steam feedwater pump system contains seven variables, as shown in Table 3. The change curves of these seven variables are shown in Figure 15.

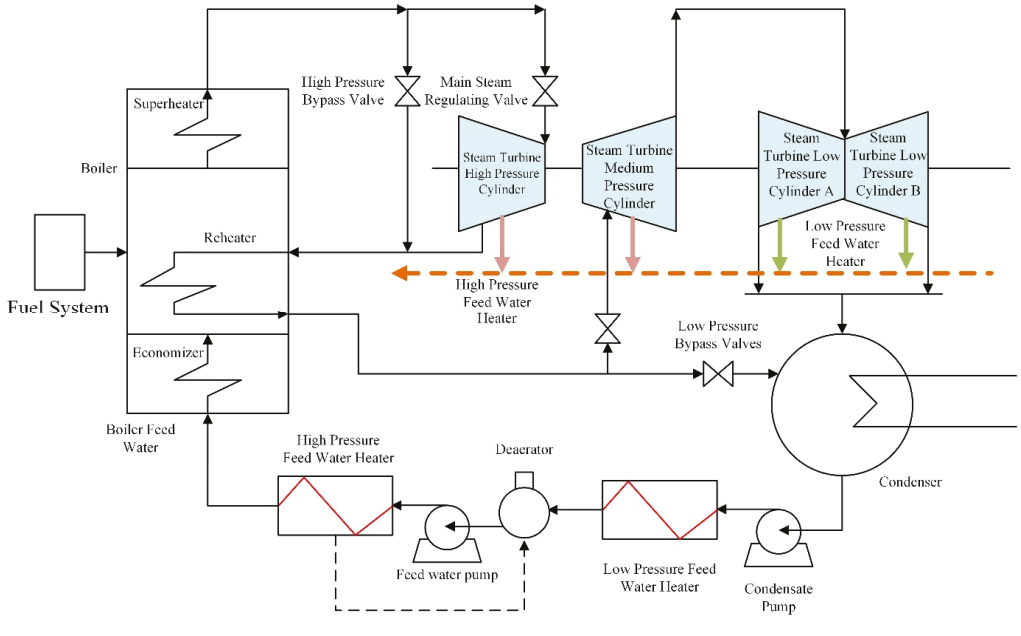
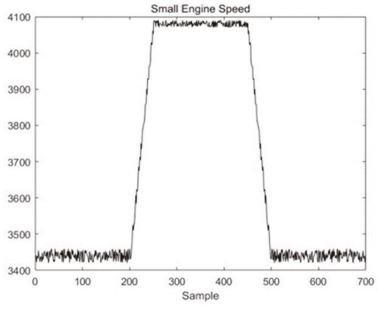
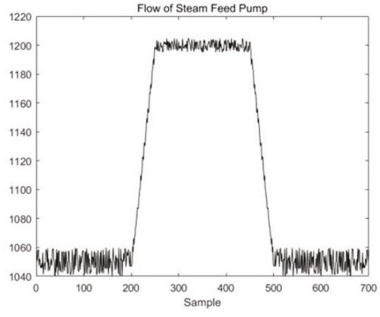


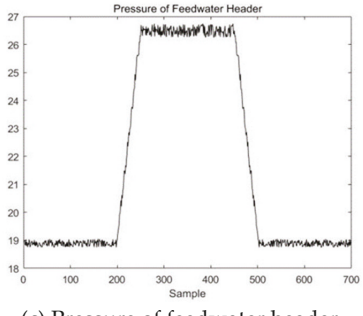
Figure 14. Schematic diagram of thermal power unit.



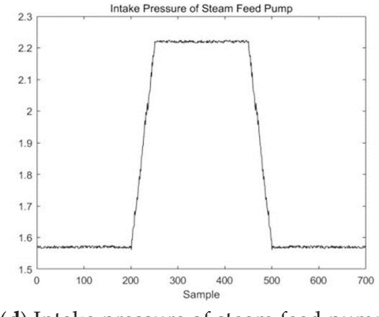
(**a**) Small engine speed



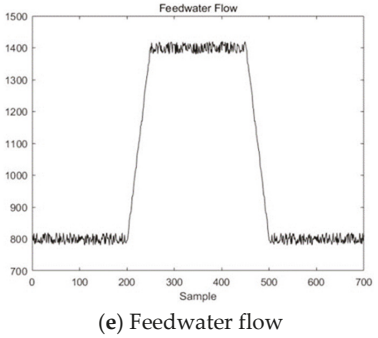
(**b**) Flow of steam feed pump



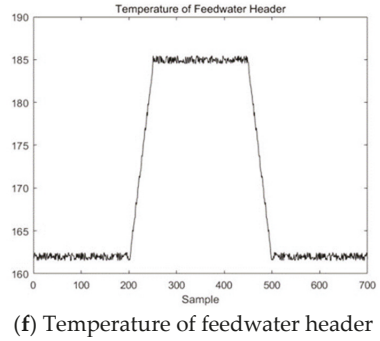
(**c**) Pressure of feedwater header



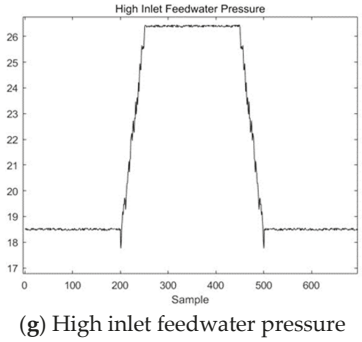
(**d**) Intake pressure of steam feed pump



(**e**) Feedwater flow



(**f**) Temperature of feedwater header



(**g**) High inlet feedwater pressure

Figure 15. The change curves of the 7 variables of the simulation data.

Table 3. Seven variables included in the steam feed pump system.

Serial Number	Variable Name
1	Small engine speed
2	The flow of the steam feed pump
3	The pressure of the feedwater header
4	Intake pressure of steam feed pump
5	Feedwater flow
6	The temperature of the feedwater header
7	High-pressure inlet feedwater pressure

First, the VLSW-MADF method is used for offline modal identification of the multimodal process. It is determined that the process begins at the 200th sampling point, enters transition mode AB from stable mode A, and then enters a new stable mode (B) after 50 sampling points. At the 450th sample point, the process starts from stable mode B, enters transition mode BA, and then enters stable mode A after 50 sample points. After mode identification, the DLPPCA method is used to model stable mode A, stable mode B, transition mode AB, and transition mode BA. The confidence limits of each mode and the matching matrices of the transition modes are saved. Since this part of the procedure is the same as the simulation of the TE process in Section 5.1, it will not be repeated here.

To prove that DLPPCA performs well with dynamic transition data, fault data in each transition mode are used for online detection; there are 400 sample points in the test dataset, and the transition from stable mode A to stable mode B begins gradually at 150th sample point. A fault signal is introduced to transition mode AB of the variable "small engine speed" to simulate a noise fault during the transition. The variable change curve is shown in Figure 16.

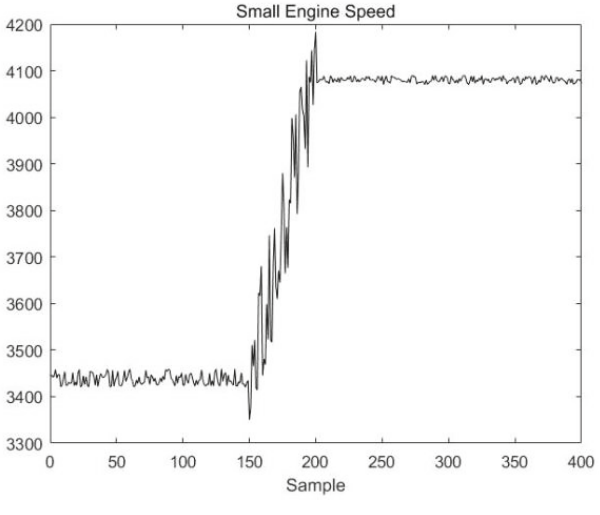


Figure 16. The change curve of the small engine speed.

Ignoring the initial mode matching process, the online monitoring effect is demonstrated for 30 consecutive sample points, starting at sample point 151. Since at the previous moment the process was in stable mode A, the online process data are monitored using stable mode A first, and the results are in Figure 17.

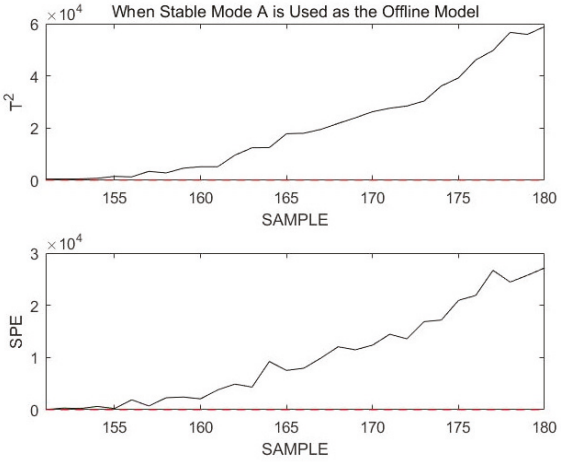


Figure 17. Monitoring 151st to 180th sample points using stable mode A as an offline model.

Obviously, all sample points exceed the confidence limit. Assuming that the current process enters a new mode since it was in stable mode A at the previous moment, the current mode may only be in transition mode AB. The online data are then remonitored using transition mode AB, as shown in Figure 18.

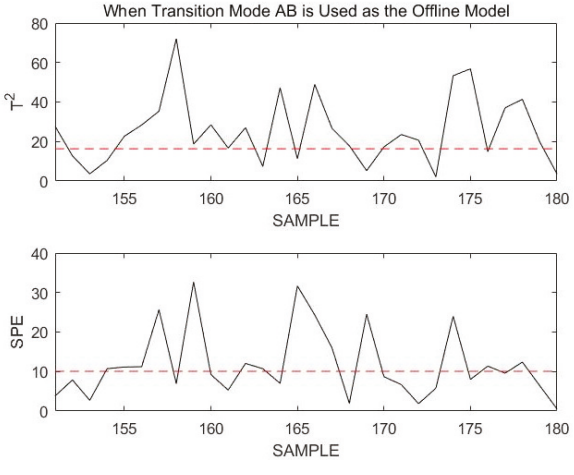


Figure 18. Monitoring 151st to 180th sample points using transition mode AB as an offline model.

If the current process mode is in transition mode AB, the statistics of the online data should be below the confidence limit when using transition mode AB for online monitoring. However, at this time, the confidence limit of the online data almost completely exceeds the confidence limit. This indicates that the online data are not in transition mode AB, but have a fault. Sample points 150 to 200 are monitored using transition mode AB, as shown in Figure 19.

From the monitoring results shown in the figure above, it can be seen that most of the sample points exceed the confidence limit when using transition mode AB to monitor the online process data. Although the sample points did not exceed the confidence limit by much, the occurrence of a fault was also identified.

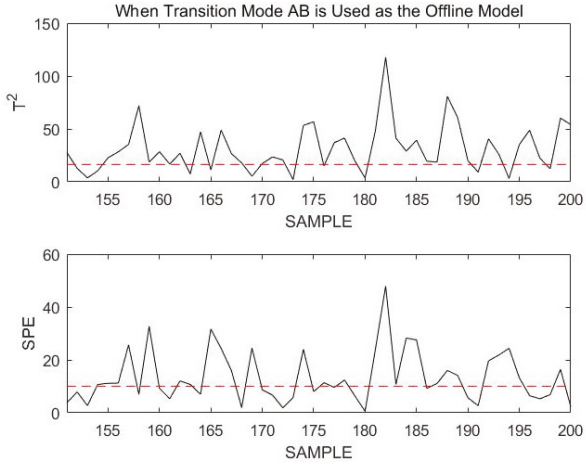


Figure 19. Monitoring 151st to 200th sample points using transition mode AB as an offline model.

These monitoring results are due to the nature of a fault itself. Transition mode data are dynamic, and the fault that occurs is that noise signals are added based on the original change trend, thereby increasing the fluctuation amplitude of the transition data. Therefore, it is difficult to monitor such faults. However, the DLPPCA method can still accurately model and monitor such transition data online. To illustrate the excellence of the DLPPCA method, a comparison is made between it, the DPCA method, and the LPPCA method without dynamic expansion. Only the modeling and monitoring methods are replaced in the comparison; all other steps remain the same.

After modeling with DPCA and monitoring the 151st to 180th sample points of the online dataset, the obtained results are shown in Figure 20.

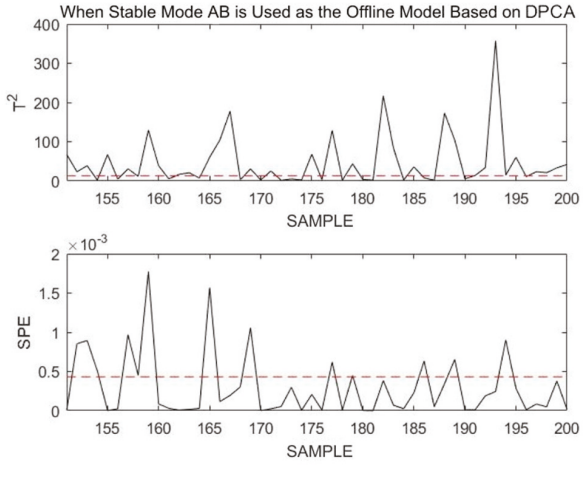


Figure 20. Monitoring 151st to 180th sample points using transition mode AB as an offline model Base on DPCA.

After modeling using LPPCA and monitoring the 151st to 180th sample points of the online dataset, the obtained results are shown in Figure 21.

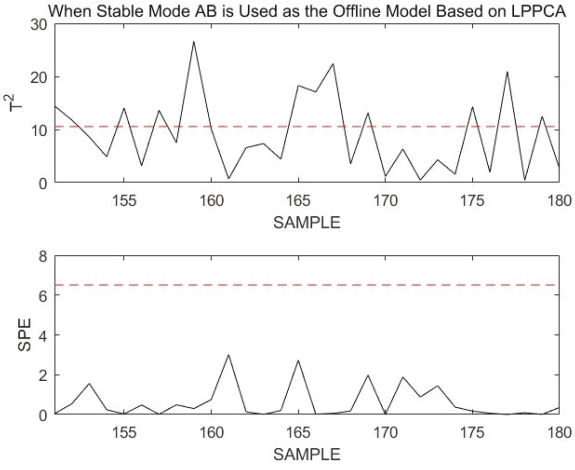


Figure 21. Monitoring 151st to 180th sample points using transition mode AB as an offline model Base on LPPCA.

Using DLPPCA, DPCA, and LPPCA, the fault detection rate (FDR) and missed alarm rate (MAR) obtained when monitoring this transitional mode failure are shown in the following Table 4.

Table 4. FDR and MAR for transitional mode failure monitoring using DLPPCA, DPCA, and LPPCA.

Methods	FDR		MAR	
	T^2	*SPE*	T^2	*SPE*
DLPPCA	76%	54%	24%	46%
DPCA	62%	26%	38%	74%
LPPCA	40%	0%	60%	100%

When the transition mode fails, the missed alarm rate (MAR) for modeling and monitoring with DPCA or LPPCA are much higher than that with DLPPCA. By comparison, the superiority of the DLPPCA method is proven. Generally, the DLPPCA method presented in this paper performs better on dynamic transition mode data and is more suitable for modeling and monitoring multimodal processes. Compared with DPCA, the accuracy of DLPPCA for fault monitoring is higher. This is because the combined use of the LPP method enables the extraction of a manifold structure that is more representative of the essential characteristics of the data while maintaining the nonlinear structure. DLPPCA fully considers both the global Euclidean structure and the local neighborhood structure of the dataset, instead of considering only one of these aspects. The false alarm rate of LPPCA for fault monitoring is much higher than that of DPCA and DLPPCA. This is due to the dynamic characteristics of autocorrelation and cross-correlation of process variables in real industrial processes. The traditional PCA-based approach is unable to extract dynamic relationships from the data, which makes it difficult to reveal the types of relationships between measured variables.

Next, another comparative simulation is used to illustrate the necessity of modeling stable modes and transition modes separately. This comparative simulation uses the DLPPCA method to perform overall offline modeling on the process containing stable mode A, stable mode B, transition mode AB, and transition mode BA without distinguishing between them.

On this basis, online data with the same trend as that of the corresponding offline data are selected for monitoring; there are 700 sample points in the online dataset. At the

200th sample point, stable mode A changes to transition mode AB, and after 50 additional sample points, stable mode B is entered. At the 450th sample point, stable mode B switches to transition mode BA and then becomes stable mode A after 50 more sample points. All online data are monitored, and the results are shown in Figure 22.

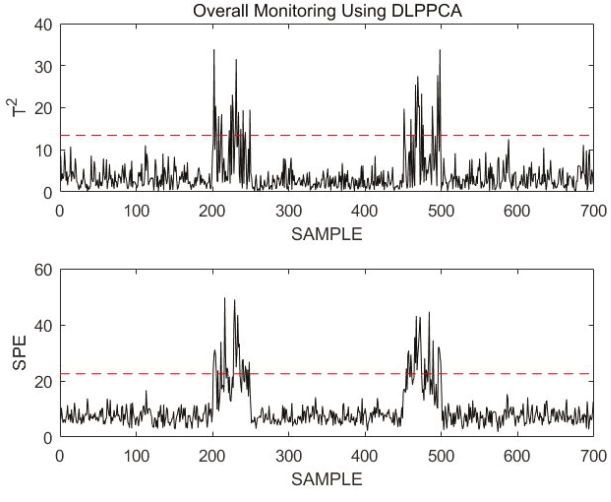


Figure 22. Overall monitoring using DLPPCA.

It was clearly seen that the normal transition mode process was incorrectly identified as a fault during the overall monitoring process. This is not the case when the stable modes and transition modes are modeled and monitored separately. The necessity of modeling stable modes and transition modes separately can be demonstrated.

Through the above simulation using an actual dataset from a power plant motor, the following can be proven:

1. The DLPPCA method is more accurate than existing methods when modeling and monitoring transition modes. Compared with DPCA and LPPCA, the DLPPCA proposed in this paper has higher modeling accuracy. For transitional mode faults that are difficult to accurately monitor with other methods, accurate results can also be obtained by using DLPPCA.
2. Modeling stable modes and transition modes separately can improve the accuracy of online monitoring. If the multimodal process is indiscriminately modeled as a whole, the normal transitional modal process can easily be mismonitored as a fault by an online monitoring approach. Modeling stable modes and transition modes separately enables us to avoid such errors and make online monitoring more accurate.

6. Conclusions

In this paper, a new multimodal process detection method is presented. In the offline phase, the VLSW-MADF test is used to identify the inherent modes, separating the stable modes from the transition modes. Then, based on the results of mode identification, the stable modes and the transition modes are modeled separately using the proposed DLPPCA method, and the confidence limit and matching matrix of each mode are saved for online mode recognition and monitoring. The VLSW-MADF test is fast and accurate in mode identification, and DLPPCA performs better on transitional mode data than traditional methods. In the online monitoring phase, this paper takes full advantage of the previous moment's historical mode and presents a new online mode identification method; this method is discussed separately and online mode identification is performed only when necessary, which reduces the computational load and improves the efficiency of mode

identification. The feasibility and efficiency of the proposed method have been evaluated through case studies involving the TE process and power plant data. Several comparisons and simulations have been made. The results show that the proposed multimodal process fault monitoring method based on the VLSW-MADF test and DLPPCA improve the efficiency and accuracy of multimodal data monitoring.

In previous research work, multiple PCA and multiple PLS are considered as the most classical methods for multimode process monitoring. Zhao [25,26] used historical data to build a single PCA or PLS model for each mode. However, this approach splits the useful information hidden between data sequences and is highly dependent on similarity measure algorithms. The proposed DLPPCA model in this paper considers the serial correlation of process data and makes full use of the global Euclidean structure and local neighborhood structure of the dataset by introducing manifold learning. The simulation results show that DLPPCA performs better than the conventional method on transition mode data. Meanwhile, the DLPPCA monitoring model can detect faults in time not only in steady mode but also in transition mode. For the problem of an unknown modeling data mode, Tan [29] uses variable-length sliding windows to extract the correlation changes of offline normal operation data and achieves the division of stable modal data and transitional modal data according to the similarity of correlation between windows. However, in the process of mode identification, the influence of the selection of the boundary parameter α on the mode identification accuracy and monitoring effect is enormous. The boundary parameter α needs to be selected by a large number of repeated experiments and expert experience. The VLSW-MADF test method proposed in this study innovatively uses the smoothness of the data as the basis for the identification of stable and transitional modes, and can accurately determine the onset of transitional modes.

Although the method in this paper achieves better results on two numerical simulation cases, there is still much room for improvement and some limitations. For example, we still perform dynamic feature extraction and analysis by constructing an augmented matrix with time lag properties. However, this method will increase the dimensionality of the data matrix and increase the computational effort. In addition, continuous learning or lifelong learning has become a key research focus in machine learning, and many researchers have introduced continuous learning into the field of process monitoring and fault diagnosis. For example, Zhang [48] investigated a single model with continuous learning capability to monitor continuous modes and achieved good results. In future research, the authors will consider improvements to existing algorithms and prefer to extend PCA to a framework of continuous learning or adaptive updating of the overall model to propose a more effective approach for industrial process monitoring.

Author Contributions: Conceptualization, S.W. and J.T.; methodology, J.T.; validation, J.T. and Y.W.; writing—review and editing, S.W., Y.W., and J.T.; visualization, Y.W.; supervision, Y.C.; project administration, S.W.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Key Research and Development Program of China (2019YFE0105000) and in part by the National Natural Science Foundation of China (61973057).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available TE process dataset was analyzed in this study. These data can be found here. [http://web.mit.edu/braatzgroup/links.html (accessed on 10 March 2020)]. The power plant data presented in this study are available upon request from the corresponding author. The data are not publicly available due to intellectual property protection.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

$\boldsymbol{A}$	Projection matrix for manifold feature extraction	T_{ti}	Matching matrix for historical transition modes
D	A $n \times n$ diagonal matrix	$\widetilde{T_{ti}}$	The first ω column vectors of T_{ti}
D_d	Distance vector between T_{online} and $\widetilde{T_{ti}}$	${T_\alpha}^2$	The confidence limit for T^2
D_{ii}	Importance of each sample	t_s	The t-statistic of the ADF test
F	Low dimensional manifold = $[f_1, f_2, \ldots, f_n]$	W	Weight matrix
H	Window length for the first stage	W_{ij}	Relationship between two samples
H^*	Smoothness matrix of the first stage	X	Sample set = $[x_m(1), x_m(2), \ldots, x_m(\mathrm{n})]^T$
I	Unit matrix	X^*	Sample set after dynamic expansion
k	Minimum number of generalized eigenvalues	$\overline{X}$	Trend variable
L	Window length for the second stage	$\widetilde{X}$	Time series = $[x_1, x_2, \ldots, x_n]^T$
L_P	A Laplacian matrix defined = $D - W$	$\overline{X}_{new}$	Sample dataset to be reidentified and tested
l	Number of time lags	$x^T(t)$	The observation vector at moment t
m	Number of variables	$\sum$	The covariance matrix of F
m_i	The matching value between T_{online} and $\widetilde{T_{ti}}$	β	Coefficient of time trend
n	Number of samples	γ_i	A trending term
P	The projection matrix when using PCA	δ	Coefficient of presenting process roots
$\mathbb{R}$	Set of real numbers	$\hat{\delta}$	The estimated value of δ
SPE_α	The confidence limit for the SPE	$\hat{\sigma}_\delta$	Standard errors
T	The feature data extracted by DLPPCA	ε_t	White noise sequence
T_{online}	The online matching matrix	η	An intercept constant called drift

References

1. Yin, S.; Li, X.; Gao, H.; Kaynak, O. Data-Based Techniques Focused on Modern Industry: An Overview. *IEEE Trans. Ind. Electron.* **2015**, *62*, 657–667. [CrossRef]
2. Ji, H.; He, X.; Shang, J.; Zhou, D. Incipient Fault Detection with Smoothing Techniques in Statistical Process Monitoring. *Control. Eng. Pract.* **2017**, *62*, 11–21. [CrossRef]
3. Peng, X.; Tang, Y.; Du, W.; Qian, F. Multimode Process Monitoring and Fault Detection: A Sparse Modeling and Dictionary Learning Method. *IEEE Trans. Ind. Electron.* **2017**, *64*, 4866–4875. [CrossRef]
4. Quiñones-Grueiro, M.; Prieto-Moreno, A.; Verde, C.; Llanes-Santiago, O. Data-Driven Monitoring of Multimode Continuous Processes: A Review. *Chemom. Intell. Lab. Syst.* **2019**, *189*, 56–71. [CrossRef]
5. Aldrich, C.; Auret, L. *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*; Springer: London, UK, 2013; ISBN 978-1-4471-5184-5.
6. Fan, J.; Wang, W.; Zhang, H. AutoEncoder Based High-Dimensional Data Fault Detection System. In Proceedings of the 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), Emden, Germany, 24–26 July 2017; pp. 1001–1006.
7. Zhao, C.; Gao, F. Fault-Relevant Principal Component Analysis (FPCA) Method for Multivariate Statistical Modeling and Process Monitoring. *Chemom. Intell. Lab. Syst.* **2014**, *133*, 1–16. [CrossRef]
8. Zhang, S.; Zhao, C. Hybrid Independent Component Analysis (H-ICA) with Simultaneous Analysis of High-Order and Second-Order Statistics for Industrial Process Monitoring. *Chemom. Intell. Lab. Syst.* **2019**, *185*, 47–58. [CrossRef]
9. Wang, J.; Zhong, B.; Zhou, J.L. Quality-Relevant Fault Monitoring Based on Locality-Preserving Partial Least-Squares Statistical Models. *Ind. Eng. Chem. Res.* **2017**, *56*, 7009–7020. [CrossRef]
10. Lee, J.-M.; Yoo, C.; Choi, S.W.; Vanrolleghem, P.A.; Lee, I.-B. Nonlinear Process Monitoring Using Kernel Principal Component Analysis. *Chem. Eng. Sci.* **2004**, *59*, 223–234. [CrossRef]
11. Ku, W.; Storer, R.H.; Georgakis, C. Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 179–196. [CrossRef]
12. Bakshi, B.R. Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE J.* **1998**, *44*, 1596–1610. [CrossRef]
13. Harrou, F.; Kadri, F.; Khadraoui, S.; Sun, Y. Ozone Measurements Monitoring Using Data-Based Approach. *Process Saf. Environ. Prot.* **2016**, *100*, 220–231. [CrossRef]
14. Roweis, S.T.; Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef]
15. Balasubramanian, M.; Schwartz, E.L. The Isomap Algorithm and Topological Stability. *Science* **2002**, *295*, 7. [CrossRef]
16. Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **2003**, *15*, 1373–1396. [CrossRef]
17. Wong, W.K.; Zhao, H.T. Supervised Optimal Locality Preserving Projection. *Pattern Recognit.* **2012**, *45*, 186–197. [CrossRef]

18. Yu, J. Local and Global Principal Component Analysis for Process Monitoring. *J. Process Control* **2012**, *22*, 1358–1373. [CrossRef]
19. Luo, L. Process Monitoring with Global–Local Preserving Projections. *Ind. Eng. Chem. Res.* **2014**, *53*, 7696–7705. [CrossRef]
20. Wu, Y.; Fu, Z.; Fei, J. Fault Diagnosis for Industrial Robots Based on a Combined Approach of Manifold Learning, Treelet Transform and Naive Bayes. *Rev. Sci. Instrum.* **2020**, *91*, 015116. [CrossRef]
21. Hwang, D.-H.; Han, C. Real-Time Monitoring for a Process with Multiple Operating Modes. *Control. Eng. Pract.* **1999**, *7*, 891–902. [CrossRef]
22. Lane, S.; Martin, E.B.; Kooijmans, R.; Morris, A.J. Performance Monitoring of a Multi-Product Semi-Batch Process. *J. Process Control* **2001**, *11*, 1–11. [CrossRef]
23. Ma, H.; Hu, Y.; Shi, H. A Novel Local Neighborhood Standardization Strategy and Its Application in Fault Detection of Multimode Processes. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 287–300. [CrossRef]
24. Ng, Y.S.; Srinivasan, R. An Adjoined Multi-Model Approach for Monitoring Batch and Transient Operations. *Comput. Chem. Eng.* **2009**, *33*, 887–902. [CrossRef]
25. Zhao, S.J.; Zhang, J.; Xu, Y.M. Monitoring of Processes with Multiple Operating Modes through Multiple Principle Component Analysis Models. *Ind. Eng. Chem. Res.* **2004**, *43*, 7025–7035. [CrossRef]
26. Zhao, S.J.; Zhang, J.; Xu, Y.M. Performance Monitoring of Processes with Multiple Operating Modes through Multiple PLS Models. *J. Process Control* **2006**, *16*, 763–772. [CrossRef]
27. Kosanovich, K.A.; Piovoso, M.J.; Dahl, K.S.; MacGregor, J.F.; Nomikos, P. Multi-Way PCA Applied to an Industrial Batch Process. In Proceedings of the 1994 American Control Conference—ACC '94, Baltimore, MD, USA, 29 June–1 July 1994.
28. Kosanovich, K.A.; Dahl, K.S.; Piovoso, M.J. Improved Process Understanding Using Multiway Principal Component Analysis. *Ind. Eng. Chem. Res.* **1996**, *35*, 138–146. [CrossRef]
29. Tan, S.; Wang, F.; Peng, J.; Chang, Y.; Wang, S. Multimode Process Monitoring Based on Mode Identification. *Ind. Eng. Chem. Res.* **2012**, *51*, 374–388. [CrossRef]
30. Marx, B.D. A User's Guide to Principal Components. *J. Am. Stat. Assoc.* **1992**, *87*, 1242. [CrossRef]
31. Xu, D.; Wang, Y. An Automated Feature Extraction and Emboli Detection System Based on the PCA and Fuzzy Sets. *Comput. Biol. Med.* **2007**, *37*, 861–871. [CrossRef]
32. He, K.; Li, X. Time–Frequency Feature Extraction of Acoustic Emission Signals in Aluminum Alloy MIG Welding Process Based on SST and PCA. *IEEE Access* **2019**, *7*, 113988–113998. [CrossRef]
33. Zhang, Y.; Wang, Z.; Zhang, J.; Ma, J. PCA Fault Feature Extraction in Complex Electric Power Systems. *AECE* **2010**, *10*, 102–107. [CrossRef]
34. Yong-dong, W.; Dong-wei, X.; Peng, P.; Yi, L.; Gui-jun, Z.; Xiao, X. Kernel PCA for Road Traffic Data Non-linear Feature Extraction. *IET Intell. Transp. Syst.* **2019**, *13*, 1291–1298. [CrossRef]
35. Li, K.; Wu, Y.; Song, S.; Sun, Y.; Wang, J.; Li, Y. A Novel Method for Spacecraft Electrical Fault Detection Based on FCM Clustering and WPSVM Classification with PCA Feature Extraction. *Proc. Inst. Mech. Eng. Part G J. Aerosp. Eng.* **2017**, *231*, 98–108. [CrossRef]
36. He, Q.; Ding, X.; Pan, Y. Machine Fault Classification Based on Local Discriminant Bases and Locality Preserving Projections. *Math. Probl. Eng.* **2014**, *2014*, 923424. [CrossRef]
37. Lv, Y.; Yuan, R.; Shi, W. Fault Diagnosis of Rotating Machinery Based on the Multiscale Local Projection Method and Diagonal Slice Spectrum. *Appl. Sci.* **2018**, *8*, 619. [CrossRef]
38. Luo, H.; Tang, Y.Y.; Li, C.; Yang, L. Local and Global Geometric Structure Preserving and Application to Hyperspectral Image Classification. *Math. Probl. Eng.* **2015**, *2015*, 917259. [CrossRef]
39. Zhang, Z.; Zhu, X.; Zhao, J.; Xu, H. Image Retrieval Based on PCA-LPP. In Proceedings of the 2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Wuxi, China, 14–17 October 2011.
40. Zhang, E.-H.; Ma, H.-B.; Lu, J.-W.; Chen, Y.-J. Gait Recognition Using Dynamic Gait Energy and PCA+LPP Method. In Proceedings of the 2009 International Conference on Machine Learning and Cybernetics, Baoding, China, 12–15 July 2009.
41. Yang, Q.; Ba, C.; Li, C.; Wu, D. An Ensemble Fault Diagnosis Approach for Multimodal Process. In Proceedings of the 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xiamen, China, 22–25 October 2017.
42. Richard, P. Modified Fast Double Sieve Bootstraps for ADF Tests. *Comput. Stat. Data Anal.* **2009**, *53*, 4490–4499. [CrossRef]
43. Worden, K.; Iakovidis, I.; Cross, E.J. New Results for the ADF Statistic in Nonstationary Signal Analysis with a View towards Structural Health Monitoring. *Mech. Syst. Signal Process.* **2021**, *146*, 106979. [CrossRef]
44. Aylar, E.; Smeekes, S.; Westerlund, J. Lag Truncation and the Local Asymptotic Distribution of the ADF Test for a Unit Root. *Stat. Pap.* **2019**, *60*, 2109–2118. [CrossRef]
45. Vosloo, J.; Uren, K.R.; van Schoor, G.; Auret, L.; Marais, H. Exergy-Based Fault Detection on the Tennessee Eastman Process. *IFAC-Pap. Line* **2020**, *53*, 13713–13720. [CrossRef]
46. Chen, D.; Li, Z.; He, Z. Research on Fault Detection of Tennessee Eastman Process Based on PCA. In Proceedings of the 2013 25th Chinese Control and Decision Conference (CCDC), Guiyang, China, 25–27 May 2013.

47. Li, H.; Xiao, D. Fault Diagnosis of Tennessee Eastman Process Using Signal Geometry Matching Technique. *EURASIP J. Adv. Signal Process.* **2011**, *2011*, 83. [CrossRef]
48. Zhang, J.; Zhou, D.; Chen, M. Monitoring Multimode Processes: A Modified PCA Algorithm with Continual Learning Ability. *J. Process Control* **2021**, *103*, 76–86. [CrossRef]

sensors

MDPI

Article

An Imbalanced Fault Diagnosis Method Based on TFFO and CNN for Rotating Machinery

Long Zhang *, Yangyuan Liu, Jianmin Zhou , Muxu Luo, Shengxin Pu and Xiaotong Yang

School of Mechatronics & Vehicle Engineering, East China Jiaotong University, Nanchang 330013, China
* Correspondence: longzh@ecjtu.edu.cn

Abstract: Deep learning-based fault diagnosis usually requires a rich supply of data, but fault samples are scarce in practice, posing a considerable challenge for existing diagnosis approaches to achieve highly accurate fault detection in real applications. This paper proposes an imbalanced fault diagnosis of rotatory machinery that combines time-frequency feature oversampling (TFFO) with a convolutional neural network (CNN). First, the sliding segmentation sampling method is employed to primarily increase the number of fault samples in the form of one-dimensional signals. Immediately after, the signals are converted into two-dimensional time-frequency feature maps by continuous wavelet transform (CWT). Subsequently, the minority samples are expanded again using the synthetic minority oversampling technique (SMOTE) to realize TFFO. After such two-fold data expansion, a balanced data set is obtained and imported to an improved 2dCNN based on the LeNet-5 to implement fault diagnosis. In order to verify the proposed method, two experiments involving single and compound faults are conducted on locomotive wheel-set bearings and a gearbox, resulting in several datasets with different imbalanced degrees and various signal-to-noise ratios. The results demonstrate the advantages of the proposed method in terms of classification accuracy and stability as well as noise robustness in imbalanced fault diagnosis, and the fault classification accuracy is over 97%.

Keywords: imbalanced data; data expansion; continuous wavelet transform; synthetic minority oversampling technique; convolution neural network

Citation: Zhang, L.; Liu, Y.; Zhou, J.; Luo, M.; Pu, S.; Yang, X. An Imbalanced Fault Diagnosis Method Based on TFFO and CNN for Rotating Machinery. *Sensors* **2022**, *22*, 8749. https://doi.org/10.3390/s22228749

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 19 October 2022
Accepted: 10 November 2022
Published: 12 November 2022

1. Introduction

Rotating machinery has been widely used as an indispensable part of industrial production [1]. The most noticeable factor in industrial production is safety [2], so the monitoring of the condition and diagnosis of malfunctions of rotating machinery have been a concern for more and more scholars [3,4]. The most common and easily damaged parts of rotating machinery are bearings and gears, which will lead to the paralysis of the entire mechanical system, property losses, and even casualties. Therefore, more advanced and universal fault diagnosis technology is urgently needed to identify faults in bearings and gears in rotating machinery so as to reduce losses [5,6].

To date, the most commonly applied methods for rotating machinery failure detection can be classified into three main groups: model-based [7], signal processing-based [8], and data-driven [9]. However, the model-based approach is challenging in establishing physical or mathematical models for relatively complex mechanical equipment [10]. Signal processing-based methods require a great deal of human knowledge to design some suitable features and understand the properties of the signals [11]. As such, these two techniques are difficult to promote in practical applications and have poor uniformity. On the contrary, the data-driven fault approach can effectively avoid the above disadvantages. It achieves bearing or gear failure classification and diagnosis by mining rules and connections within big data [12,13].

Deep learning, represented by convolutional neural networks (CNN), is a typical data-driven fault diagnosis method that enables end-to-end fault diagnosis without prior

knowledge [14]. At present, researchers have applied CNNs in fault diagnosis of rotating machinery. For instance, Janssens et al. proposed a feature learning model for condition monitoring based on CNN [15]. Yao et al. used an acoustic approach and CNN based on a multiscale dialog learning structure and attention mechanisms for gear fault diagnosis [16]. Zhang et al. implemented bearing fault diagnosis under different operating loads using DCNN with original signals [17].

Although the work mentioned has obtained great diagnostic results, an issue remains to be addressed: CNN-based intelligent bearing fault diagnosis algorithms often require large samples for training. Nevertheless, obtaining enough fault samples in practical applications is difficult and even impossible, so the amount of data is usually imbalanced. This small and imbalanced data will considerably affect the accuracy of the fault diagnosis model.

In practical cases, rotating machinery has been in routine operation for a long time, and faults seldom happen during the machinery work. Consequently, faulty samples are more difficult to collect than normal samples, which results in the number of faulty samples will be much smaller than the number of normal samples [18]. Small and imbalanced data (S&I data) is a common situation faced by intelligent diagnosis models [19,20]. This situation is prone to cause model overfitting resulting in poor classification results, especially for deep learning fault diagnosis [21]. Thus, the diagnosis technique is more effective in classification when the amount of data is adequate, and the various types are balanced. For example, a mass of jobs conducted by other authors obtained promising results in the case of the Case Western Reserve University bearing dataset, which is typically a database of a sufficient and balanced number of samples [22–24]. Unfortunately, the scarcity of failure samples has permeated every aspect of our lives, such as in aerospace applications where rotating devices are replaced regularly, making it almost impossible to obtain failure samples, resulting in an extreme imbalance between the different categories. Therefore, sample augments and enhancement are the research focus.

The current mainstream sample expansion techniques, such as generative adversarial networks (GAN) [25], recurrent neural networks (RNN) [26], and variational auto-encoder (VAE) [27], have been widely applied. The above three mainstream methods have the potential to augment samples for specific problems. However, deep networks often require much time to train the model and are weak in generality. In addition, as a typical data expansion method, the synthetic minority oversampling technique (SMOTE) is able to solve the problem of data imbalance, which generates new samples between two adjacent samples by linear interpolation [28]. It compensates for the drawback that random oversampling inclines to cause overfitting. Han et al. adopted the Borderline-SMOTE method to oversample with a few class boundaries in the primary data [29]. Safe-Level-SMOTE multiplies the original few class instances by different weighting factors to construct safe regions [30]. The ADASYN algorithm adaptively adjusts the weights of different minority classes in the raw dataset [31]. The application of the above SMOTE algorithm directly performs sample generation on the original data. Nevertheless, the quality of the synthesized new samples largely rests with the original samples and their neighboring representatives. It is impossible to avoid suffering from the interference of noisy components and causing a shift in the data distribution, which will significantly affect the accuracy of the subsequent diagnosis.

Short-term Fourier transform (STFT) and continuous wavelet transform (CWT), as time-frequency analysis methods, can demonstrate the characteristic changes of the signal in the two-dimensional time-frequency spectrums and have better noise suppression [32,33]. As a result, the CWT and STFT are widely used for rotating machinery fault diagnosis. For example, Chikkerur et al. presented feature enhancement on fingerprint signals based on STFT [34]. Alexakos et al. achieved STFT denoising on motor-bearing image data [35]. Kankar et al. present a bearing fault diagnosis methodology using CWT, which consists of six different base wavelets [8]. However, the CWT is superior in extracting time-frequency features compared with STFT. The STFT adopts a fixed window function. When the win-

dow function is determined, its shape will not change, and the resolution of the STFT will be determined, resulting in its sampling interval cannot decrease with increasing frequency. In contrast, the wavelet transform has an adjustable time-frequency window [36], which can visually show the change in frequency components over time and accurately analyze the scale and resolution of periodic or transient signals. In addition, the CWT is the capability to detect weak defect signals from non-stationary data, even in strong noises [37,38]. Numerous researchers have adopted generative adversarial networks (GAN) to significantly expand the CWT-denoised image data to achieve better diagnostic results [39–41]. Nevertheless, GAN requires more cost to adjust the network structure to generate better samples and suffers from the problem that the model is not generalized. Compared with the SMOTE, the GAN algorithm requires more time to expand data and suffers from poor generalization.

Based on the above analysis, this paper chooses CWT as a tool for denoising and analyzing time-frequency features. In addition, the SMOTE was employed for sample expansion, thus proposing a new imbalance data augment the model with a time-frequency feature oversampling method (TFFO). Finally, CNN is established to realize the imbalance fault diagnosis of rotating machinery. The contributions of the research are listed as follows:

1. The proposed method performs a comprehensive data expansion from different dimensions. On the one hand, the sliding segmentation method partially expands some numbers of time-domain fault samples. On the other hand, SMOTE is applied to build a balanced dataset by expanding the minority fault samples in the time-frequency images.
2. CWT is employed as a pre-processing tool to construct 2-dimensional time-frequency images and denoise the data to enhance the stability of the features. In addition, an improved CNN based on LeNet-5 is established to extract the features and automatically recognize the fault location.
3. Compared with existing mainstream data augmentation techniques such as GAN and LSTM, the TFFO-CNN-based model has better performance in the diagnosis of bearing and gear failures under two imbalanced datasets, even under the interference of noisy environments.

The remainder of this paper is organized as follows: the introduction of SMOTE, CWT, and CNN in Section 2. Section 3 presents the general idea of the imbalance fault diagnosis model. In Section 4, two experimental studies are developed to evaluate the proposed approach for determining rotating machinery faults compared to other existing approaches. Finally, conclusions and future work are provided in Section 5.

2. Methodology

2.1. Data Expansion Based on Sliding Segmentation and SMOTE

2.1.1. Sliding Segmentation

In actual practice, the machine is usually not allowed to run for long periods when a bearing or gear fails, resulting in a minimal number of vibrational fault signals that can be collected. Hence, finding a way to expand the limited signal is significant.

A sliding segmentation is employed for repeated sampling during the first data augmentation in this paper, which exploits the periodic nature of the fault signal to expand the sample. The process of selecting and moving the sliding window is as follows:

1. Window size. Theoretically, the size of the essential sliding window should be greater than or equal to one rotation period. Therefore, according to the rotation speed and the sampling frequency, the number of sample points produced by a rotation period of the bearing or gear can be calculated, that is, the minimum length of the sliding window.
2. Sliding step. The most basic principle for choosing the moving step size is that it should be less than one rotation period and that the step size should be smaller than the sliding window size. On the one hand, when the sliding step is small, the overlap rate of adjacent samples is higher, and the difference of expanded samples is slight,

which is easy to cause overfitting of training. On the contrary, when the sliding step size is more extensive, due to the limitation of sample length, the expanded sample size is smaller, which is easy to cause training underfitting.

3. Starting point and sliding direction. In general, the first point of the raw data is set as the starting point of the sliding window on the premise that the data are correct. Until the last point of the data, the sliding direction should move in the direction of time.

As depicted in Figure 1, Assuming that the sample length is N, the slip window size is W, the moving step size is B, and the number of samples after sliding segmentation is M, it can be expressed as:

$$M = \frac{N - W + B}{B} \tag{1}$$

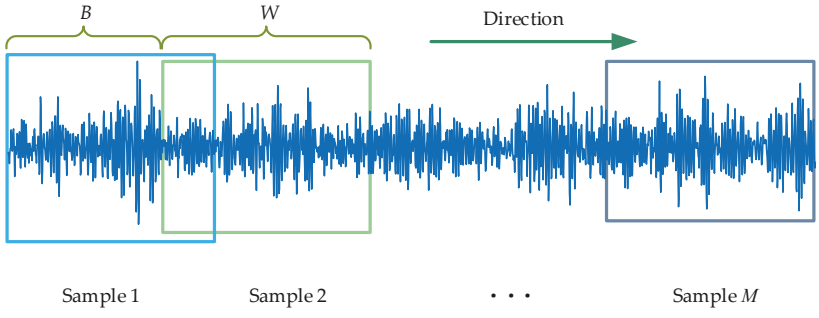


Figure 1. Illustration of the sliding segmentation. It mainly contains four key factors, including window size, sliding step and starting point, and sliding direction.

2.1.2. Introduction to SMOTE

The SMOTE is an improved scheme based on the random oversampling algorithm [28], as shown in Figure 2. The essential concept is to analyze the minority samples and add new samples to the data set. The approximate flow of the algorithm is based on the K nearest neighbor sample points of each sample point. It randomly selects N adjacent points to multiply the difference by a threshold in the range of (0,1) to achieve the purpose of synthesis of data. The process of the SMOTE algorithm is as follows:

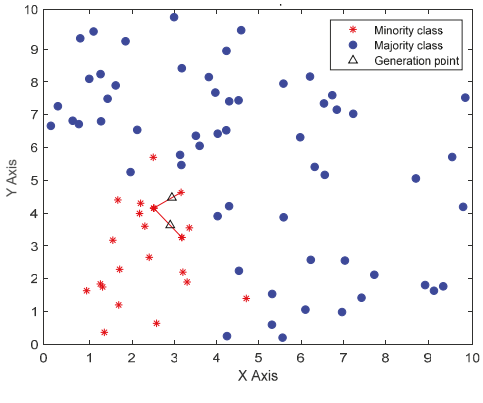


Figure 2. Illustration of SMOTE algorithm. The blue balls, red asterisks, and black triangles, respectively represent the majority classes, the minority classes, and the generation points.

1. For each minority category X_0, its distance from all surrounding samples is calculated on the basis of the Euclidean distance, and K nearest neighbor is obtained.

2. According to the sample imbalance ratio, the sampling ratio is set. For each minority sample, several samples are randomly selected from their K nearest neighbors.
3. For each randomly selected nearest-neighbor sample, create a new random point on the line segment connecting the pattern and the selected neighbor, as follows:

$$X_{new} = X_0 + w(X - X_0) \tag{2}$$

where w is a uniform random variable in the range (0,1), X_{new} is the generated point, X_0 is the minority category, X is the surrounding sample.

2.2. Introduction of CWT

2.2.1. Wavelet Transform

Compared to the 1-dimensional time-domain signal, the 2-dimensional time-frequency domain matrix has more information as an image and can represent a more complex structure [42]. The one-dimensional time domain signal is converted into a two-dimensional characteristic spectrum by CWT in this paper. The CWT has excellent local description ability in the time and frequency domains [43]. Its temporal resolution and frequency resolution change with scale, which are in accordance with the characteristics of slow variations of the low-frequency signal and rapid variations of the high-frequency signal. CWT overcomes the shortcomings of the short-time Fourier transformation and continues its idea of time-frequency analysis of signals [44]. It is an excellent time-frequency analysis technique for transient analysis [45]. In fact, the bearing and gear fault signals contain many transient shock components [46]. Therefore, CWT has a unique advantage in dealing with rotating machinery failure datasets.

When the vibration signal:

$$x(t) \in L^2(R) \tag{3}$$

Then the wavelet transform $w_{wt}(a, b)$ can be expressed as:

$$w_{wt}(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\ \psi_{a,b}(\frac{t-b}{a}) dt \tag{4}$$

where $\psi_{a,b}$ is a family of wavelet functions. It can be obtained from $\psi(t)$.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi(\frac{t-b}{a}) \tag{5}$$

where a is the translation factor, and b is the scale parameter. $a, b \in R, a > 0$. In this paper, the size of b is set as the length of each sample.

2.2.2. Selection of the Wavelet Basis Function

The selection of a wavelet basis function depends on the nature of the signal being analyzed and the purpose of the application. Among the existing wavelet functions, the Morlet wavelet has the form of an exponential attenuation vibration, which is very similar to the shock vibration response caused by bearing faults [47], so the Morlet wavelet has been widely studied in rolling bearing resonance demodulation technology.

The Morlet wavelet basis function is composed of a complex trigonometric function multiplied by an exponential attenuation function, and the expression is as follows:

$$\psi(t) = e^{-\frac{t^2}{2}} e^{j\omega_0 t} \tag{6}$$

After stretching and translating, it can be expressed as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} e^{-\frac{(t-b)^2}{2a^2}} e^{j\omega_0 \frac{(t-b)}{a}} \tag{7}$$

The acquisition of time-frequency images will be described in Section 4.

Following repeated sampling and expansion of some samples by sliding segmentation, the CWT is adopted to decompose the vibration signal of each sample into a wavelet coefficient matrix. The time-frequency distribution can characterize the joint information between the time and frequency domains and highlight the relationship between the signal and the operating state of the equipment. After the above processing, the signal benefits the model training and recognition.

2.3. Improved CNN Model Construction

CNN has been developed rapidly in recent years and has become an efficient method for feature recognition [48]. CNN is composed of multiple convolutional, pooling, and fully connected layers, whose architecture is displayed in Figure 3. The structure of the CNN established in this paper is designed based on the LeNet-5 network [49]. The essence of CNN is to build a filter that can extract many different features of the input data. The output of the previous layer is used as the input of the next layer, and compelling feature extraction is achieved layer by layer.

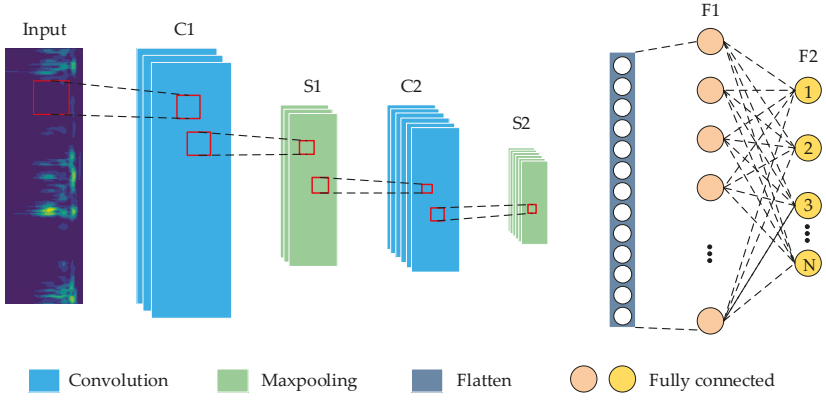


Figure 3. The architecture of LeNet-5-based CNN. It mainly contains two multiple convolutional, two pooling layers, and two fully connected layers. The time-frequency images are input to the first convolutional layer, and the classification of the output layer is achieved by the softmax function.

In Figure 3, two convolution kernels of different sizes are constructed to extract the image's main features and fine local features, respectively. The upper layer feature maps are convolved, and the Rectified Linear Unit (ReLU) activation function obtains the new feature maps. ReLU, as the most common nonlinear activation function in neural networks, can effectively improve the nonlinear fitting ability of neural networks [50], as shown in Equation (8). The Max-pooling layer uses the most significant local features to reduce the dimensionality of the feature input and compress the number of parameters after the convolution layer. The fully connected layer connects all features of the previous layer, integrates local information with the classification of the convolutional or pooling layer, and sends the output values after Sigmoid activation to the classifier. Sigmoid is a smooth and continuous activation function, also known as a logistic function, which can map a real number to the interval of (0,1) [51]. It is shown in Equation (9). The Sigmoid and ReLU activation functions are shown in Figure 4. Dropout is introduced to improve the model's generalization ability and prevent overfitting [52]. The dropout algorithm randomly hides some units with a probability of failure during the training process [53]. Finally, the error loss between the predicted and actual values of the labels is calculated using a binary cross-entropy loss function for backpropagation, which has the ability to adjust the offsets in each layer to minimize the loss function.

$$f(x) = \max\{0, x\} \tag{8}$$

$$g(x) = \frac{1}{1 + e^{-x}} \tag{9}$$

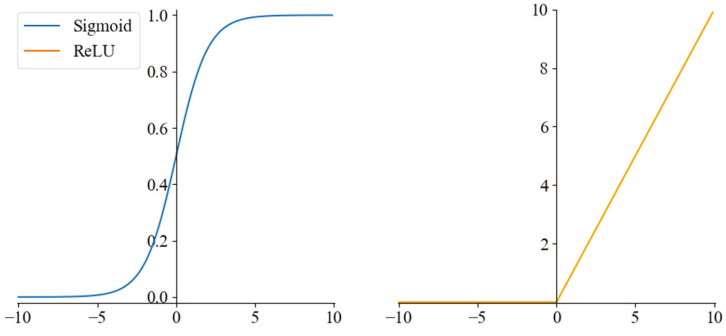


Figure 4. The Sigmoid and ReLU activation function.

Compared with the LeNet-5 network [49], the specific improvements of the improved CNN model in this paper are as follows:

(1) The LeNet-5 network uses a fixed 5 × 5 convolutional kernel, but the convolutional kernel is too large to extract the fine local features in the sample. In this paper, two convolution kernels of different sizes are constructed to extract the image's main features and fine local features, respectively.
(2) To enhance the robustness of the model, the improved model adds a ReLU activation function after the convolution layer, which is useful to avoid gradient saturation and reduce the training time.
(3) The LeNet-5 network uses two fully connected layers, which is computationally intensive and time-consuming. Therefore, in the improved CNN in this paper, only one fully connected layer is used after the convolution module with the Softmax layer for output;
(4) A Dropout technique is added before the fully connected layer. This approach reduces the degree of correlation between neurons, thus avoiding network overfitting and improving the generalizability of the model.

3. Proposed Approach

Aiming at the problem of reduced accuracy of model diagnosis due to S&I data, this paper proposes a new approach for imbalanced fault diagnosis of rotating machinery based on TFFO and CNN. Figure 5 shows the flowchart of the imbalanced fault diagnostic process, including the collection of acceleration signals and faulty signals expanded by sliding segmentation, the time-frequency feature extraction of the one-dimensional signals using CWT, the minority samples are balanced through SMOTE, illustration of CNN model, and visualization of the classification result. The main steps are described as follows:

1. Data acquisition. Bearings or gears experimental objects with different types of failure are loaded using different test benches. Acceleration sensors are installed to collect and construct vibration signal datasets.
2. First data expansion. On the basis of the above vibration signal dataset, slip segmentation sampling is performed to extend the range of samples. Moreover, CWT is applied to denoise and generate time-frequency maps containing rich information in time and frequency domains.
3. Second data augment. Samples from a few classes are analyzed to create new samples among the randomly selected nearest neighbor samples using SMOTE. The sampling rate is set according to the data imbalance rate to balance the time-frequency map dataset.

4. Diagnostic model. The time-frequency map dataset is fed into a designed CNN model comprising convolution, pooling, and fully connected layers with Softmax to output gear and bearing fault diagnosis results.
5. Visualization. The model output is visualized using the T-SNE algorithm and the confusion matrix.

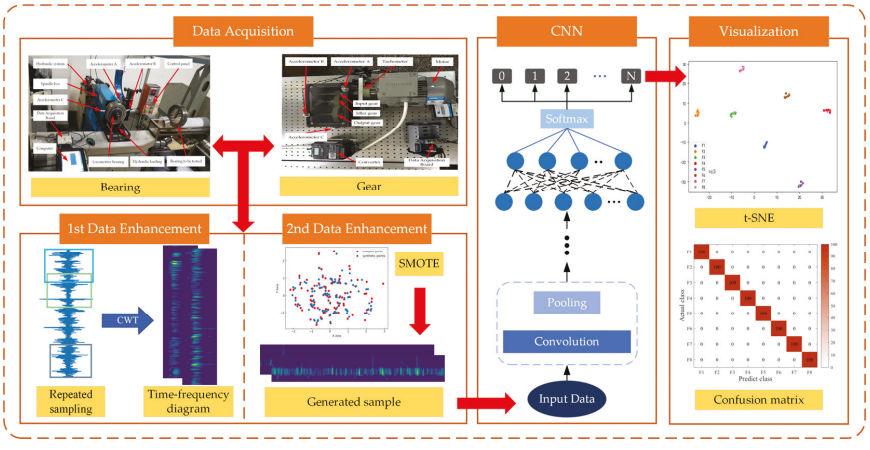


Figure 5. Imbalanced fault diagnosis flow chart of rotating machinery based on TFFO and CNN. First, the bearing and gearbox raw vibration signals are collected. Second, sliding segmentation is used for repeated sampling, and CWT is applied to generate time−frequency images. Third, the SMOTE is utilized to generate minority samples again. Finally, an improved CNN based on LeNet−5 is established to achieve intelligent fault diagnosis while the features are visualized by t−SNE, and results are displayed by a confusion matrix.

4. Experiments and Results

In this section, experimental studies are conducted on bearing and gear, respectively: one is the locomotive bearing dataset, and the other is the public gearbox dataset from Zhejiang University. Meanwhile, the latest data expansion approaches are used for comparisons, such as GAN and LSTM. Moreover, the CNN model learning conditions and the diagnosis accuracy also deserve our attention. We apply t-SNE to project the features of each layer into a two-dimensional representation, which better describes the layer-by-layer learning capability of the CNN network model. The fault diagnosis results are quantified in detail by a multi-classification confusion matrix, and related charts will comprehensively demonstrate the fault recognition accuracy.

It is worth noting that this paper aims to simulate a realistic situation with a small number of fault samples, which provides a new idea for the imbalance fault real-time diagnosis of rotating machinery. Therefore, the model should use as few real fault samples as possible during the experiment. The author used only individual sensor data to construct the imbalance dataset in this paper's bearing and gear fault diagnosis experiments.

4.1. Case Study 1: The Locomotive Bearing Dataset

4.1.1. Experimental Setup

The bearing data is employed from a locomotive depot of the China Railway Administration. The data set of bearing faults are real faults, not artificial processing faults. The current locomotive bearing dynamic detection system model of the Railway Bureau is the JL-501 series. The main body of the bearing detection system consists of the bearing test rig and the software detection device, as shown in Figure 6. The locomotive wheelset bearing is driven and loaded with the detection platform in this paper. The spindle speed is set at 500 rpm, and the radial load is 1.4 MPa. The locomotive bearings used in the experiment

are NJ2232WB series cylindrical roller bearings with an outer diameter of 290 mm and an inner diameter of 160 mm. Vibration signals are obtained by three model CA-YD-187T accelerometers fixed at the outer ring of the bearings and a Ni-USB-4431 acquisition card. The sampling frequency is 20 kHz. Eight types of locomotive bearing failures, including normal state, are shown in Table 1, and the corresponding locomotive bearings are shown in Figure 7.

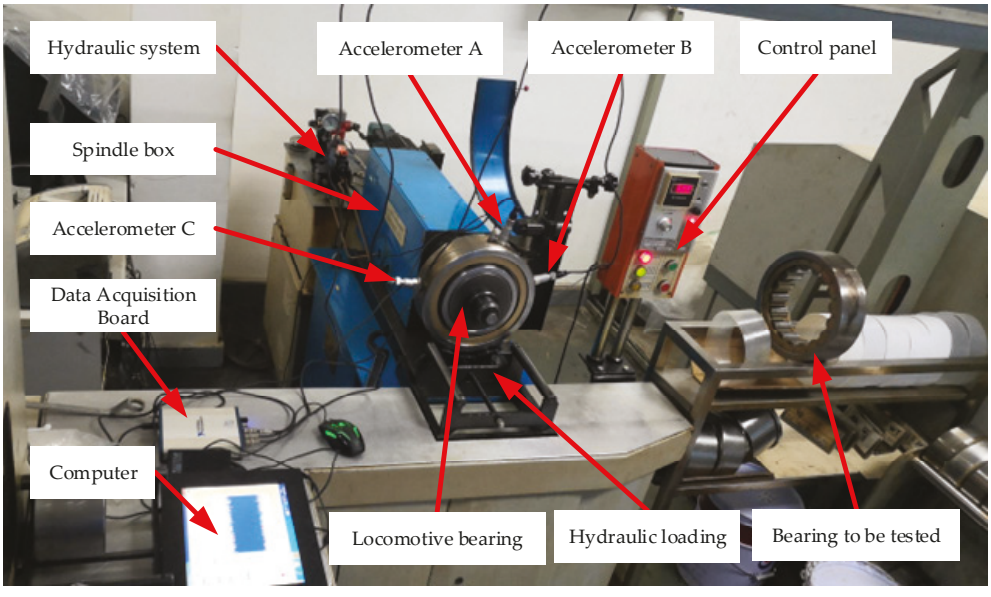


Figure 6. The locomotive bearing test rig. It is from a locomotive depot of the China Railway Administration. It mainly contains a hydraulic system, a spindle box, hydraulic loading, and three accelerometers at different locations.

Table 1. A detailed description of the bearing data set.

Label	Fault Type	Length	Original Samples	Dataset 1	Dataset 2	Dataset 3
F1	Slight failure of cage	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F2	Compound failure of cage and rolling body	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F3	Slight failure of rolling body	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F4	Slight failure of inner ring	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F5	Severe failure of inner ring	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F6	Slight failure of outer ring	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F7	Severe failure of outer ring	102400	42 × 2400	50 × 2400	50 × 2400	50 × 2400
F8	Normal	1200000	500 × 2400	50 × 2400	250 × 2400	500 × 2400

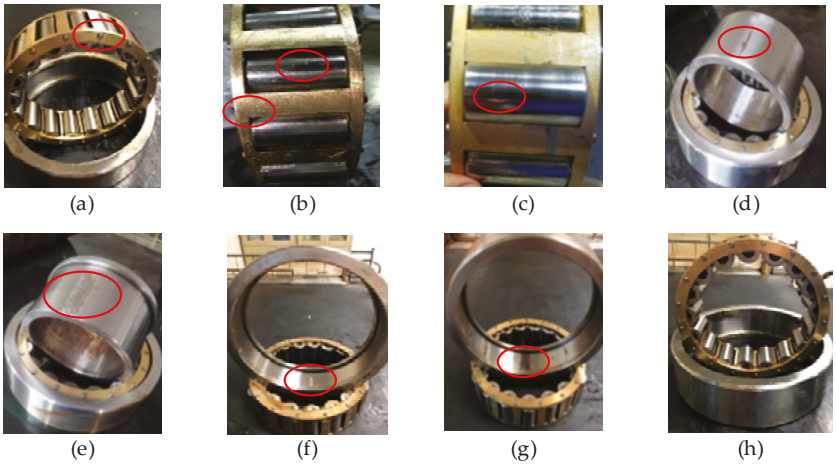


Figure 7. Different types of defective bearings: (**a**) F1; (**b**) F2; (**c**) F3; (**d**) F4; (**e**) F5; (**f**) F6; (**g**) F7; (**h**) F8. The red circle in the figure indicates the location of the defect.

4.1.2. Preprocessing of Data and Parameter Selection

For the bearing data set of 8 categories, Figure 8 shows the corresponding time-domain signals. There are 1,200,000 data points for healthy bearings and 102,400 for the other seven types of fault data. According to the sampling frequency of 20 kHz and the speed of 500 rpm, the sample length of this experiment is 2400. Thus, this bearing data set has about 42 faulty samples and about 500 normal samples.

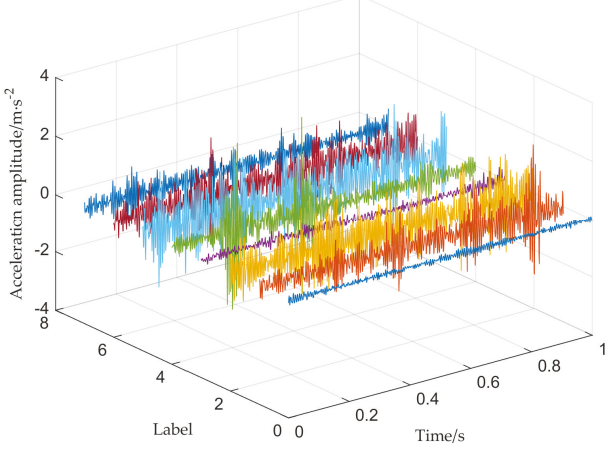


Figure 8. Time domain signal of F1—F8. It mainly contains eight types of fault signals in Table 1.

Three different imbalanced levels datasets are constructed artificially based on the number of normal bearing samples, where Dataset 1 has 50 normal samples, Dataset 2 has 250, and Dataset 3 has 500. The imbalance ratios for the normal and faulty samples of the three datasets are 1, 0.2, and 0.1, respectively. The specific process of building the three datasets is as follows:

In addition to the normal bearing samples F8, the remaining seven types of fault samples were expanded to build a balanced data set. In the first data expansion using the sliding segmentation method, the window size was 2400, and the moving step size

was 2000. The number of repetition points was 400. Finally, the original samples were expanded to form Dataset 1. In the second data augment using the TFFO, the sample size was increased by different multiples for the Dataset 2 and Dataset 3 of the different imbalance ratios in Table 1. Ultimately, the number of samples for each category remained consistent with the number of samples for the bearings in the healthy state.

One-dimensional time-domain signals are transformed into time-frequency feature images using CWT, where the scale factor is set to 2400, depending on the length of each sample. The frequency range of the vertical axis in the time-frequency diagram indicates the fault resonance frequency range (2.5 kHz–5 kHz), which is determined by the fault itself. For example, the resonant frequency of the bearing refers to the fact that the bearing rotation will cause a shock at the fault location, and this shock will produce the phenomenon of inherent frequency resonance. Figure 9 shows the time-frequency images of the original and generated samples after the transformation by CWT. The differentiation between the various types of samples is still evident in Figure 9. We can see that the fault feature information is mainly distributed in the middle frequency band (2.5 kHz–5 kHz), and the generated time-frequency image is similar to the primitive image under the same health state.

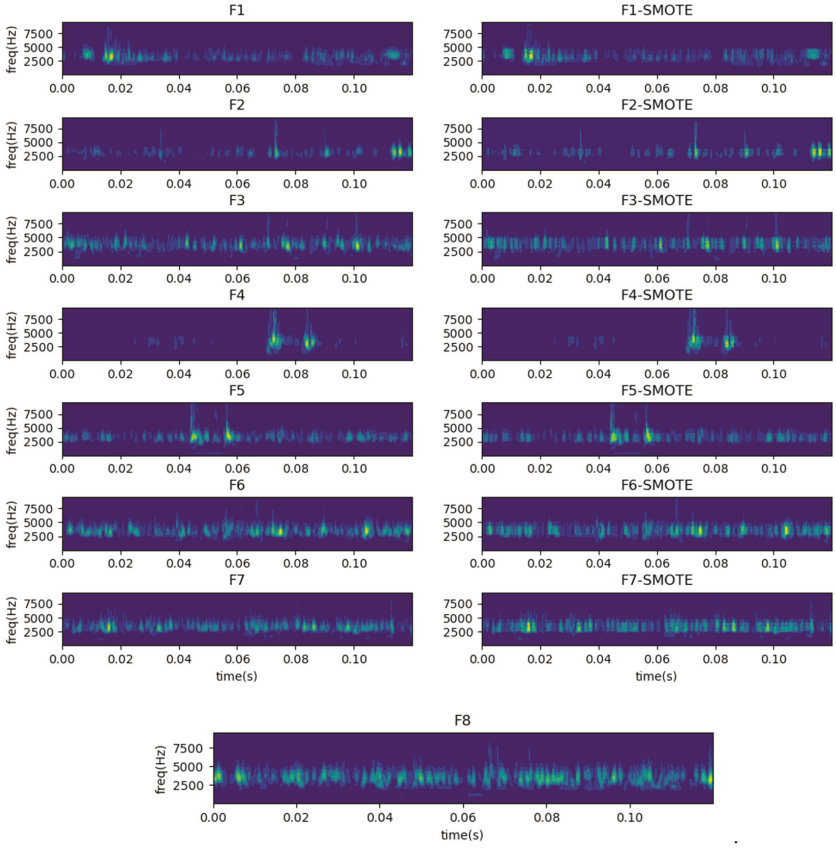


Figure 9. Time-frequency images of the original samples and generated samples. It mainly contains a healthy-bearing sample and seven fault-bearing samples, and seven generated samples.

There is no need to make the generated time-frequency sample utterly consistent with the original ones. The identical samples are meaningless in the training process of the model. Fortunately, the vibration signals of bearings and gears are distinctly periodic. Thus,

the model can perform highly accurate fault diagnosis when the generated samples contain comprehensive fault information. In addition, many studies took flip, rotate, and randomly crop as image data augment tools to make different samples [53–55].

TFFO is a method of oversampling based on feature space, in which a new sample is formed by synthesizing new characteristics between a primitive sample and the nearest neighbor. The distribution of the data generated by TFFO technology is very similar to the original data, which causes the generated and the original picture to be challenging to distinguish and recognize by human eyes. However, this is not difficult for CNN.

After secondary data expansions, the class-balanced dataset was divided into three parts: 60% for training, 20% for validation, and 20% for testing. What needs to be emphasized is that the test set data is fixed and does not contain any generated samples, while the training and validation sets are randomly assigned from the remaining samples in proportion to the remaining samples. Subsequently, the data are input into 2-dimensional CNN for fault location identification. In order tto reduce the effect oerrors; e errors, ten random experiments are passed to maximize accuracy and minimize loss of validation set data. The trained model is then employed to classify the data from the test set. The choice of hyperparameters in the CNN model significantly influences the accuracy of subsequent fault diagnosis. In this paper, the epochs, batch size, learning rate, and dropout were 60, 50, 0.001, and 0.5, respectively. The structure and parameters of CNN are described in Table 2.

Table 2. The detailed structure of CNN.

Layer	Kernel	Strides	Output Size	Activation	Padding	Param
Input	/	/	$98 \times 2400 \times 1$	/	/	0
C1	4×4	4	$24 \times 600 \times 64$	ReLU	Valid	1088
S1	2×2	2	$12 \times 300 \times 64$	/	/	0
C2	2×2	2	$6 \times 150 \times 128$	ReLU	Valid	32,896
S2	2×2	2	$3 \times 75 \times 128$	/	/	0
F1	128	/	128	Sigmoid	/	3,686,528
F2	N	/	N	Softmax	/	1032

The software and hardware facilities used for data processing in this experiment are as follows: Win10 64-bit operating system, AMD Ryzen 7 3800X 8-Core processor, 32 GB running memory, a program running Python3.6, Spyder, Tensorflow1.13.1.

4.1.3. Diagnosis Results and Visualization

Figure 10 shows the loss and accuracy curves after balancing Dataset 1, Dataset 2, and Dataset 3 using the proposed TFFO and CNN methods. In all datasets, the loss value decreases to about 0.01, and the accuracy rate reaches 100% when the iteration reaches the 40th round. From the 40th round onward, the model further converges until it is stable. We can clearly learn that the model has promising diagnostic results and strong generalization performance.

A multiclassification confusion matrix is introduced to conduct a detailed quantitative analysis of fault diagnosis results, which provides a comprehensive view of the types and the specific number of misclassifications of the actual fault types. Figure 11 visually represents the classification of the test set after sample balancing for the three data sets in Table 1. Figure 11a shows the classification results for the test set in Dataset 1. There are ten samples for each fault type, and the categories F5, F7, and F8 are misclassified with a misclassification rate of 7.5%. The imbalance ratio of Dataset 2 is 5 to 1 in Figure 11b. After the dataset is balanced, the sample size increases significantly, and the misclassification phenomenon is much improved than in Dataset 1. From Figure 11c, it can be observed that the result is satisfactory under Dataset 3. The final accuracy reaches 100%, although the Dataset 3 sample ratio reaches 10 to 1, and the imbalance is very high.

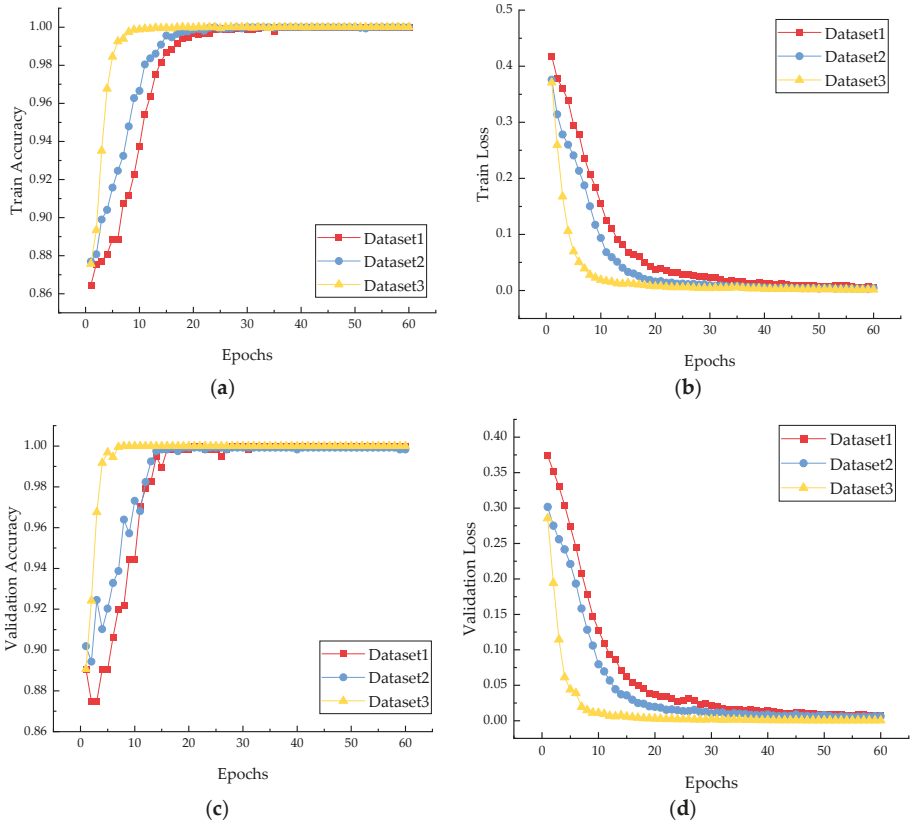


Figure 10. Experimental results for balanced bearing Dataset 1, Dataset 2, and Dataset 3: (**a**) train accuracy; (**b**) train loss; (**c**) validation accuracy; (**d**) validation loss.

Figures 10 and 11 explain more intuitively the effect of class imbalance on the final classification accuracy, which shows that the balance between different data types significantly affects the final accuracy. The loss curve shows that the model converges faster after secondary data expansion. The accuracy curves and confusion matrix results show that the model is more stable and more accurate after data balancing.

T-SNE (T-distributed stochastic neighbor embedding) algorithm is a nonlinear manifold learning algorithm to visualize high-dimensional data [56]. The algorithm aims to keep the neighborhood distribution characteristics of high-dimensional data and low-dimensional data consistent as much as possible. The KL divergence is used to measure the difference between two distributions, and the gradient descent method is used to minimize the distribution difference.

T-SNE dimension reduction was performed on two convolution layers and a fully connected layer to visualize the model effect in 2dCNN. As can be seen from Figure 12a,b, the distribution among the eight classes of samples is disordered and covers significantly. It is impossible to distinguish the types of faults. However, the situation gradually improves as the number of layers in the network increases.

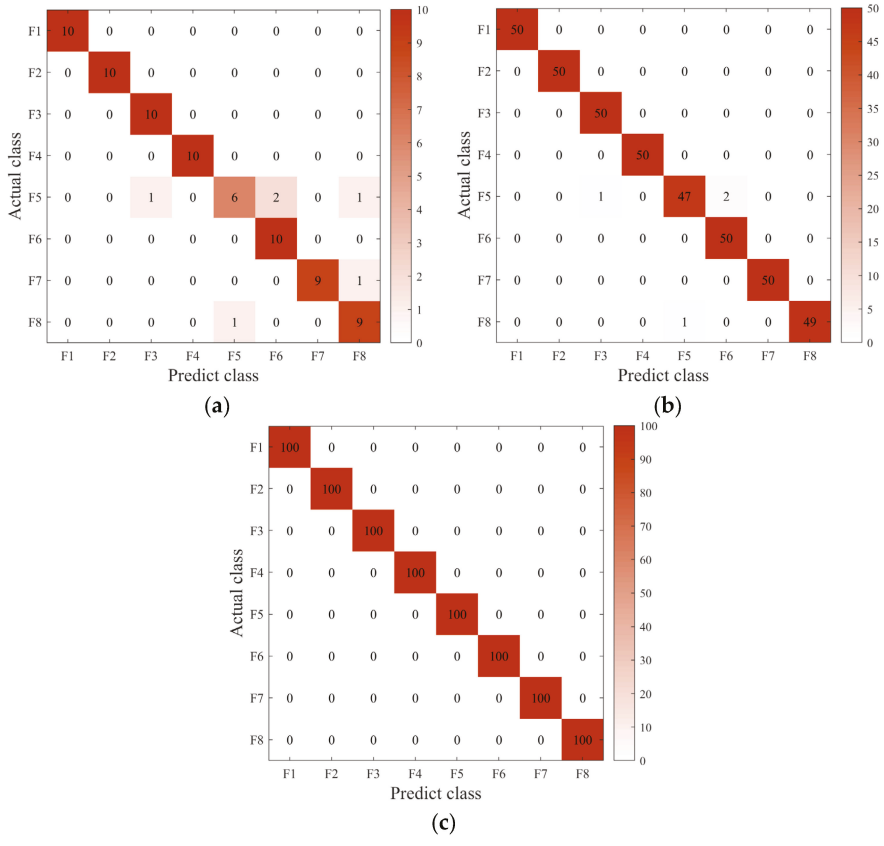


Figure 11. The confusion matrix under different datasets: (**a**) Dataset 1; (**b**) Dataset 2; (**c**) Dataset 3.

Figure 12c shows the sample distribution of the last fully connected layer. There is a clear distinction between different types of faults and no misclassification. Nevertheless, the original imbalanced data input with the same parameters and network structure of 2d-CNN, its full-connection layer classification effect is still not ideal. The label F8 in Figure 12d is the normal sample. It is impossible to minimize the intraclass distance due to the large proportion of imbalances leading to a more dispersed distribution. Several samples labeled F5 were mistakenly assigned to other areas, leaving some scattered and accessible.

In addition, we constructed a series of experiments to compare and analyze the proposed model. First, we show the average accuracy of the proposed approach ten times under different imbalance ratios and noise levels.

Gaussian white noise is added to the original signal to generate noisy signals with different signal-to-noise ratios (SNRs) to simulate the industrial environment. Then different imbalance ratio data sets are constructed and inputted into the proposed model for data augment. All experiments were expanded on the original scaled data using TFFO until the categories were balanced. Each group of experiments calculated the average accuracy of 10 tests and the extreme range between the maximum and minimum accuracy. The average accuracy rate reflects the accuracy of the model. When the value is more significant, the model is more accurate. Moreover, the extreme range indicates the degree of generalization of the model, and the smaller the value, the better the generalization effect. It is worth noting that the amount of bearing data in the healthy state is much more considerable than in the faulty state. Hence, the imbalance ratio can reach 10 to 1.

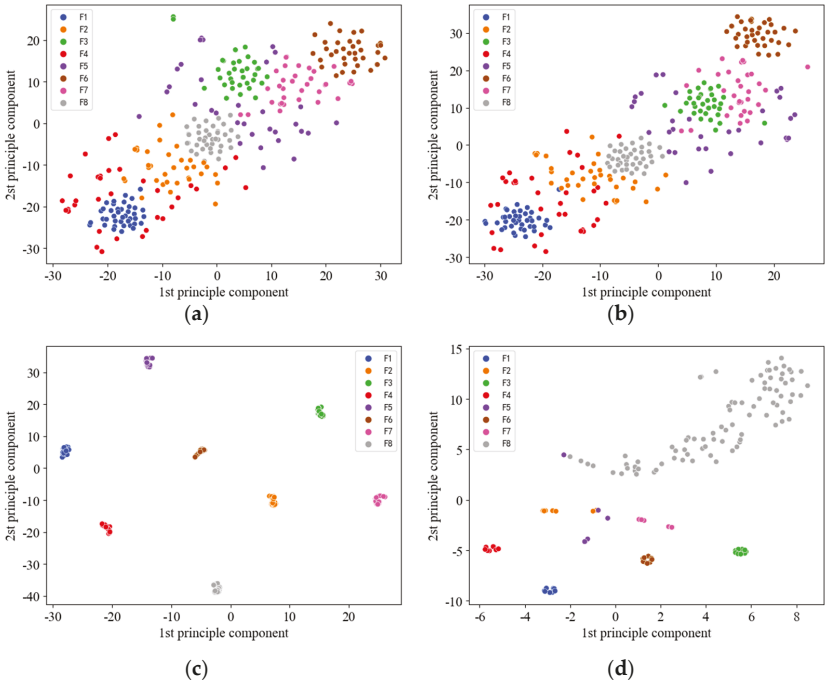


Figure 12. The visualization by t−SNE of the learned features in the Conv2D layer and Fully connected layer of Dataset 3: (**a**) layer C1 in the balanced Dataset 3; (**b**) layer C2 in the balanced Dataset 3; (**c**) layer F1 in the balanced Dataset 3; (**d**) layer F1 in the imbalanced Dataset 3.

Table 3 shows the test accuracy of the proposed method in different SNRs. We can learn that after using the proposed TFFO for data augment, the performance of 2dCNN in classifying imbalanced data has been significantly enhanced, and the test accuracy of Dataset 2 and Dataset 3 reaches 97% and 99% or more, respectively. On the contrary, the test accuracy of Dataset 1 is between 91% and 98%, which is not a satisfactory result. The number of expanded samples increases as the imbalance ratio continues to increase. Subsequently, the average accuracy at an arbitrary SNR is increasing. A satisfactory result of 100% accuracy was achieved using Dataset 3 in a 0 dB noise environment.

Table 3. Comparison of the performance comparison of different SNRs.

Dataset	Judging Criteria/%	−4 dB	−2 dB	0 dB	2 dB	4 dB
Dataset 1	Average accuracy	91.38	93.625	98.75	93	95.5
	Max-Min	6.25	8.75	2.5	2.5	2.5
Dataset 2	Average accuracy	97.75	97.15	99.35	98.3	98.8
	Max-Min	0.5	1.25	1	1.25	0.75
Dataset 3	Average accuracy	99.275	99.6	100	99.65	99.325
	Max-Min	0.75	0.5	0	0.25	0.5

Through the analysis of the experimental results, it is easy to find that TFFO and 2dCNN can overcome the data imbalance problem well. On the other hand, we show the 10-fold average diagnostic accuracy of different methods at different noise levels using Dataset 3. In this section, in order to validate the proposed imbalance fault diagnosis model, the proposed method was compared with two mainstream data enhancement algorithms: GAN [57] and LSTM [58]. The two prevalent networks are broadly described as follows:

The generators and discriminators in the GAN have constantly been adversarial and improved [59,60]. Random input noise is eventually converted into a signal similar to the target output. Different classes of faulty samples are inputted into the GAN until the number of faulty samples equals that of normal samples.

LSTM is an improved network based on recurrent neural networks. It can predict the next data point based on the correlation of the temporal signal. The process is repeated until a fault signal with the same length as the normal signal is generated. In this paper, the structure of LSTM is 1000-32-32-1, the Dropout is 0.2, and the batch size is 16. Adam is selected as the optimizer.

Figure 13 shows the variation of the 10-test accuracy for the four methods at five SNRs. Figure 14 shows the box plot based on the accuracy of ten times. The proposed TFFO and CNN-based imbalance fault diagnosis approach have more than 99% accuracy at different SNRs. In contrast, the test accuracies of model CWT-GAN-CNN and model LSTM-CNN in a −4 dB noise environment are only about 95% and 93%, respectively. The diagnostic performance of the GAN and LSTM networks is approximately the same at each SNR but slightly lower than the TFFO. CWT-CNN method has the most significant variance in accuracy values at each SNR, and the model is the most unstable. It is difficult for CWT-CNN models to identify fault types when the data set is severely imbalanced. In a word, the TFFO-CNN approach shows optimal performance in terms of accuracy and stability.

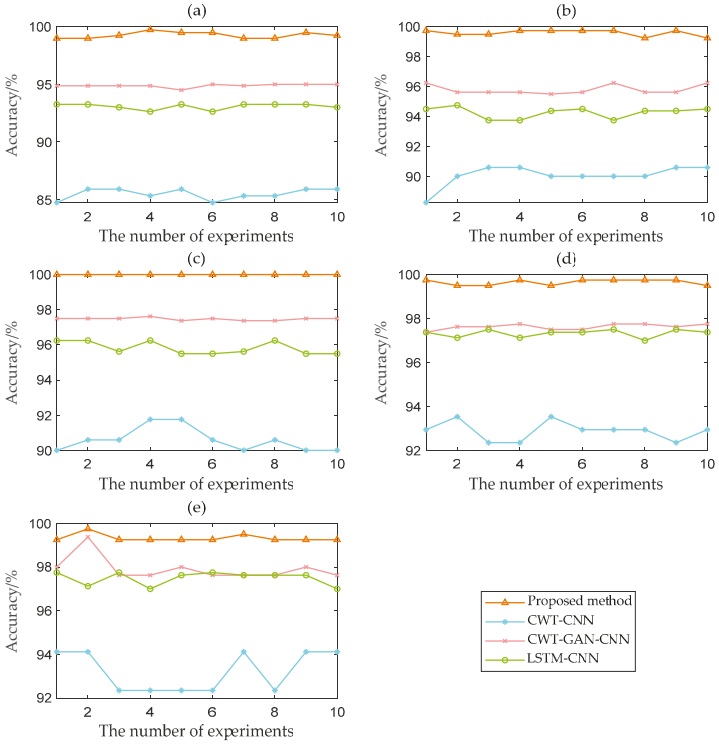


Figure 13. Accuracy curves of four models with different SNRs: (**a**) SNR = −4 dB; (**b**) SNR = −2 dB; (**c**) SNR = 0 dB; (**d**) SNR = 2 dB; (**e**) SNR = 4 dB.

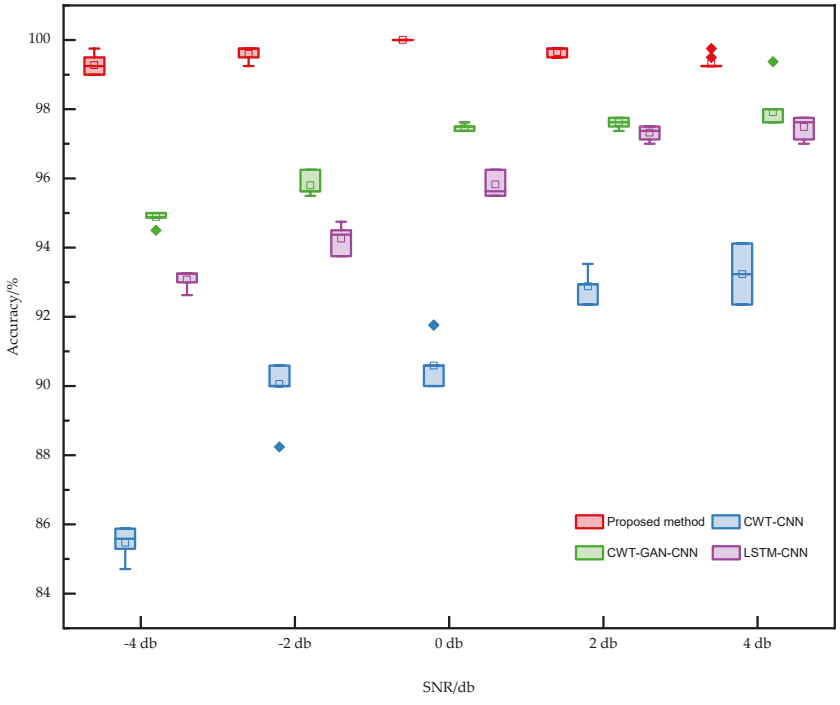


Figure 14. Comparison of the performance of different models with different SNRs. It mainly contains four models, including the proposed method, CWT−CNN, CWT−GAN−CNN, and LSTM−CNN.

For example, the data length of 12,000 points is used to expand the sample, the proposed TFFO method takes about 5 min, the GAN model takes 100 min, and the LSTM model takes 70 min. This is because the TFFO data augmentation method generates new samples by oversampling the time-frequency features. GAN and LSTM, on the other hand, require continuous training and refinement of the minority sample. Thereby, the approach proposed is also much better than other data-enhancement methods in terms of timeliness.

The hyperparameters in the proposed CNN are the optimal values of multiple artificial experiments. To further explore the effect of hyperparameters on the classification results, we perform experimental analyses on different combinations of three parameters of batch size, learning rate, and dropout using Dataset 3. As we can see from Experiment 1 in Table 4, the diagnostic accuracy reaches 100% when the batch size is over 50. However, when the batch size is too large, the model requires more epochs for training and a higher RAM. Therefore, the model is optimal when the batch size is 50. Meanwhile, when the learning rate is 0.001, the result is optimal from Experiment 2. We can see that the value of dropout has no effect on the diagnostic results using Dataset 3 from Experiment 3. However, the model is prone to overfitting when the amount of data is small. In fact, the dropout technique can effectively solve the model overfitting phenomenon.

Further, experiments are conducted on Dataset 1, which has smaller data, and the diagnostic results are shown in Table 5. Different dropout values have a significant impact on the diagnostic accuracy of the model, and the result reaches the optimum when the dropout is 0.5. Generally, the value of dropout is set to 0.5, which is a reasonable approximation to taking the geometric mean of the predictive distributions produced by the exponentially-many dropout networks [53].

Table 4. Experiments for selection of optimal parameters using Dataset3.

Experiments	Initial Conditions	Variants	Test 1	Test 2	Test 3	Test 4	Test 5
1	Learning rate = 0.01 Dropout = 0.5	Batch size Accuracy	30 98.4%	40 99.1%	50 100%	60 100%	70 100%
2	Batch size = 50 Dropout = 0.5	Learning rate Accuracy	0.0001 99.2%	0.001 100%	0.01 97.9%	0.1 13.4%	1 12.5%
3	Batch size = 50 Learning rate = 0.01	Dropout Accuracy	0 100%	0.3 100%	0.5 100%	0.7 100%	0.9 100%

Table 5. Experiments for selection of optimal dropout using Dataset 1.

Initial Conditions	Variants	Test 1	Test 2	Test 3	Test 4	Test 5
Batch size = 50 Learning rate = 0.01	Dropout	0	0.3	0.5	0.7	0.9
	Accuracy	97.2%	98.67%	100%	97.9%	69.4%

In summary, when the model is optimal, the batch size, learning rate, and dropout are 50, 0.001, and 0.5 in this paper, respectively.

4.2. Case Study 2: The Gearbox Dataset

4.2.1. Experimental Setup

In this experiment, the gearbox dataset from Zhejiang University is used [61], and Figure 15 shows the experimental gear rig, which comprises an AC motor, coupling, and a two-stage gearbox. The driving power of the motor is 0.75 kW, and the maximum speed can reach 3000 r/min. The frequency converter controls the speed of the vehicle. The experiment uses three single-axis accelerometers fixed at the gearbox's input, output, and mounting plates to collect vibration signals at different locations. The number of teeth of the input, inert, and output gears is 32, 64, and 96, respectively. However, the gear may have a missing tooth, broken teeth, a crack in the tooth root, and gluing and peeling of the tooth surface. Table 4 provides a detailed description of the ten health conditions. The sampling frequency is 25.6 kHz, and the rotating speed is 2700 rpm during the experiment.

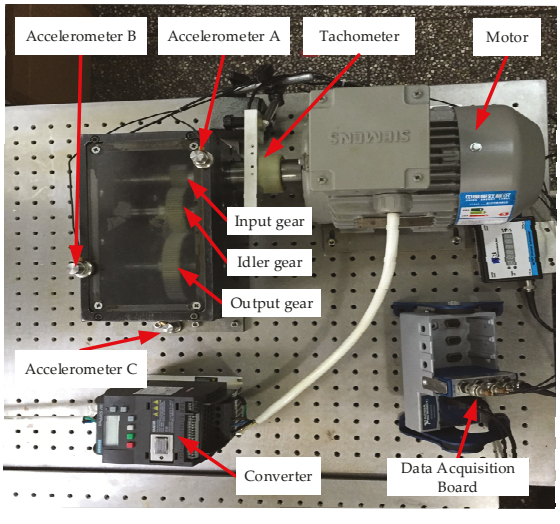


Figure 15. The gear test rig, which is from Zhejiang University and primarily contains a motor, three gears, and three accelerometers, and a data acquisition board.

4.2.2. Experimental Results

For the 10-classified gear data set, there are 240,000 data points for health status and 50,400 for each of the other nine types of fault status, and the proportion of class imbalance is about 4.76. The rotation of the gear with the maximum number of teeth according to the sampling frequency and rotational speed will produce about 569 points in one cycle. The sample length of this paper is 1200. Table 6 shows the sample changes before and after secondary data expansion.

Table 6. Introduction to gear data sets.

Label	Fault Type and Condition	Samples	Second Enhancement
C1	a broken tooth on the input gear	42 × 1200	200 × 1200
C2	a pitted tooth on the input gear	42 × 1200	200 × 1200
C3	a pitted tooth on the idler gear	42 × 1200	200 × 1200
C4	a pitted tooth and broken tooth on the output gear	42 × 1200	200 × 1200
C5	a missing tooth on the output gear	42 × 1200	200 × 1200
C6	a cracked tooth on the input gear	42 × 1200	200 × 1200
C7	a cracked tooth on the idler gear	42 × 1200	200 × 1200
C8	a cracked tooth on the output gear	42 × 1200	200 × 1200
C9	a broken tooth on the input gear and a pitted tooth on the idler gear	42 × 1200	200 × 1200
C10	normal	200 × 1200	/

This article adopts four performance indicators, accuracy, precision, recall and F1-score to indicate diagnosis ability with test data, as shown in Table 7. A higher value means better fault diagnosis performance. The CWT-CNN method is applied as a comparison method using an unbalanced dataset, while the remaining two methods use different sample expansions. Compared to the other three methods, the method proposed in this paper improved accuracy by 18.35%, 2.47%, and 7.19%, respectively. The precision increased by 19.72%, 2.39%, and 7.17%, respectively. The recall rate increased by 17.48%, 2.67%, and 6.73%, respectively. The improvement in F1-score is 18.83%, 2.53% and 6.77%, respectively. In the comparative analysis of the above data, it can be seen that the proposed approach outperforms the other three methods in all metrics, which indicates that TFFO-CNN has excellent diagnostic performance.

Table 7. Evaluation indicators for different models.

Criteria/%	Proposed Method	CWT-CNN	CWT-GAN-CNN	LSTM-CNN
Accuracy	99.50 ± 0.25	81.15 ± 1.54	97.03 ± 1.16	92.31 ± 1.54
Precision	99.25 ± 0.50	79.53 ± 0.89	96.86 ± 0.24	92.08 ± 0.78
Recall	98.71 ± 0.30	81.23 ± 0.93	96.04 ± 1.03	91.98 ± 0.34
F1-score	98.79 ± 0.29	79.96 ± 1.08	96.26 ± 0.51	92.02 ± 0.33

4.3. Discussion

This paper proposes an imbalanced fault diagnosis method based on time-frequency feature oversampling and CNN for rotating machinery. First, this paper adopts the first expansion of the fault data from the sliding segmentation method. Subsequently, the sample performs feature enhancement and denoising by the TFFO method. Finally, CNN completes the fault identification of the balanced dataset. In the analysis, three imbalanced scale datasets are constructed to verify the diagnostic performance of the model. The bearing data set is the actual operational failures of the wheelset bearings. It is challenging for researchers to obtain the fault data, but they are significant for applying diagnostic models under realistic operating conditions. Meanwhile, the robustness of the model is examined under different SNRs. The experiments were compared with three methods, CWT-CNN, CWT-GAN-CNN, and LSTM-CNN. Ten times diagnostic accuracy and box plot results show that the proposed approach outperforms the other methods in accuracy

and stability in all cases. The proposed approach takes less time to obtain higher diagnostic accuracy when processing image data. The reason is that the TFFO method is a feature-based oversampling method that is more time sensitive. Four comprehensive evaluation metrics of the laboratory artificially faulty gear dataset were extracted, indicating that the proposed method still has a high fault identification capability when dealing with other diagnostic objects and imbalanced ratios. In fact, the data expansion method proposed in this paper is not limited to the imbalance ratio in the text. It can be applied to other fault diagnosis tasks with imbalanced data sets.

5. Conclusions

This paper focuses on the imbalanced fault diagnosis problem and proposes a TFFO-CNN-based model characterized by the development of a time-frequency feature oversampling technique to reconstruct robust class balance data and further feature extraction and fault classification using the 2dCNN model. This combination gives full play to the advantages of each model. The main conclusions are summarized as follows:

(1) The proposed model constructs balanced datasets by simultaneously extending the time-domain signal and time-frequency domain features, which performs a comprehensive data expansion from different dimensions.
(2) Applying the CWT to convert vibration signals into image data allows the signal to achieve denoising and automatic feature extraction. SMOTE oversampling method is performed on the denoised time-frequency features to generate high-quality samples, which solves the problem that the other sample expansion methods do not consider the noise and result in the low quality of the generated data, such as GAN and LSTM. The time-frequency feature oversampling method that combined CWT and SMOTE can significantly reduce the sample generation time.
(3) The proposed imbalance fault diagnosis model solves the problem of inadequate model training effectively under a variety of imbalanced radios. The proposed imbalance fault diagnosis approach has more than 99% accuracy at different SNRs using bearing dataset 3. Meanwhile, compared to the other methods, including CWT-CNN, CWT-GAN-CNN, and LSTM-CNN, the method proposed in this paper improved accuracy by 18.35%, 2.47%, and 7.19% in the gear dataset, respectively. Experiments prove that the final fault recognition rate of the imbalance fault diagnosis model of rotating machinery based on TFFO, and CNN is the best among the models tested.

This approach provides a solution for imbalanced fault diagnosis of rotating machinery and demonstrates the potential of combining the time-frequency feature oversampling technique with the CNN model in fault diagnosis. While good results have been obtained using the proposed method based on bearing and gearbox datasets, further discussion is still necessary on the failure of rotating machinery where interrupted shaft failures and rotor failures, etc., also often occur. We will evaluate the proposed method with the rotor datasets in future work. Moreover, the study will further examine the performance of the proposed method for the case of compound fault diagnosis in the actual industry.

Author Contributions: Conceptualization, L.Z.; Funding acquisition, L.Z. and J.Z.; Methodology, L.Z., Y.L. and J.Z.; Software, Y.L. and X.Y.; Supervision, L.Z. and J.Z.; Validation, Y.L., M.L. and S.P.; Visualization, S.P. and X.Y.; Writing—original draft, M.L.; Writing—review & editing, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation of China (No. 51865010), the Natural Science Foundation of Jiangxi Province (No. 20212BAB204007), the Jiangxi Province Graduate Student Innovation Project (No. YC2021-S422), and the Science Research Project of the Education Department of Jiangxi Province (No. GJJ200616).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the authors upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

TFFO	Time-Frequency Feature Oversampling Technique
CNN	Convolution Neural Networks
CWT	Continuous Wavelet Transform
GAN	Generating Adversarial Networks
RNN	Recurrent Neural Networks
VAE	Variational Auto-Encoder
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WT	Wavelet Transform
SNR	Signal-to-Noise Ratios
LSTM	Long Short-Term Memory Network

Mathematical Notations

$M = \frac{N-W+B}{B}$	M is the number of samples after sliding segmentation N is the sample length W is the slip window size B is the moving step size
$X_{new} = X_0 + w(X - X_0)$	X_{new} is the generated point X_0 is the minority category X is the surrounding sample w is the uniform random variable in the range (0,1)
$x(t) \in L^2(R)$	$x(t)$ is the vibration signal $L^2(R)$ is the Hilbert Space
$w_{wt}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\ \psi_{a,b}(\frac{t-b}{a})dt$	a is the translation factor b is the scale parameter $\psi_{a,b}$ is a family of wavelet functions $w_{wt}(a,b)$ is the wavelet transform

References

1. Sharma, S.; Tiwari, S.K.; Singh, S. Integrated approach based on flexible analytical wavelet transform and permutation entropy for fault detection in rotary machines. *Meas. J. Int. Meas. Confed.* **2021**, *169*, 108389. [CrossRef]
2. Zhang, L.; Zhang, J.; Peng, Y.; Lin, J. Intra-Domain Transfer Learning for Fault Diagnosis with Small Samples. *Appl. Sci.* **2022**, *12*, 7032. [CrossRef]
3. Chen, Y.; Zhang, T.; Zhao, W.; Luo, Z.; Lin, H. Rotating Machinery Fault Diagnosis Based on Improved Multiscale Amplitude-Aware Permutation Entropy and Multiclass Relevance Vector Machine. *Sensors* **2019**, *19*, 4542. [CrossRef] [PubMed]
4. Vashishtha, G.; Kumar, R. Feature Selection Based on Gaussian Ant Lion Optimizer for Fault Identification in Centrifugal Pump. In *Recent Advances in Machines and Mechanisms*; Springer: Singapore, 2023; pp. 295–310.
5. Zhao, Z.; Li, T.; Wu, J.; Sun, C.; Wang, S.; Yan, R.; Chen, X. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Trans.* **2020**, *107*, 224–255. [CrossRef]
6. Yan, R.; Shen, F.; Sun, C.; Chen, X. Knowledge Transfer for Rotary Machine Fault Diagnosis. *IEEE Sens. J.* **2020**, *20*, 8374–8393. [CrossRef]
7. Tyagi, S.; Panigrahi, S.K. Transient Analysis of Ball Bearing Fault Simulation using Finite Element Method. *J. Inst. Eng. (India) Ser. C* **2014**, *95*, 309–318. [CrossRef]
8. Kankar, P.K.; Sharma, S.C.; Harsha, S.P. Fault diagnosis of ball bearings using continuous wavelet transform. *Appl. Soft Comput.* **2011**, *11*, 2300–2312. [CrossRef]
9. Zhang, L.; Lin, J.; Shao, H.; Zhang, Z.; Yan, X.; Long, J. End-to-end unsupervised fault detection using a flow-based model. *Reliab. Eng. Syst. Safe* **2021**, *215*, 107805. [CrossRef]
10. Di, L.; Lin, Z. Control of a flexible rotor active magnetic bearing test rig: A characteristic model based all-coefficient adaptive control approach. *Control Theory Technol.* **2014**, *12*, 1–12. [CrossRef]
11. Zhou, Q.; Yan, P.; Liu, H.; Xin, Y. A hybrid fault diagnosis method for mechanical components based on ontology and signal analysis. *J. Intell. Manuf.* **2019**, *30*, 1693–1715. [CrossRef]

12. Li, M.; Yu, D.; Chen, Z.; Xiahou, K.; Ji, T.; Wu, Q.H. A Data-Driven Residual-Based Method for Fault Diagnosis and Isolation in Wind Turbines. *IEEE Trans. Sustain. Energ.* **2019**, *10*, 895–904. [CrossRef]
13. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method. *IEEE Trans. Ind. Electron.* **2018**, *65*, 5990–5998. [CrossRef]
14. Cerrada, M.; Sánchez, R.; Li, C.; Pacheco, F.; Cabrera, D.; Valente De Oliveira, J.; Vásquez, R.E. A review on data-driven fault severity assessment in rolling bearings. *Mech. Syst. Signal Process.* **2018**, *99*, 169–196. [CrossRef]
15. Janssens, O.; Slavkovikj, V. Convolutional Neural Network Based Fault Detection for Rotating Machinery. *J. Sound Vib.* **2016**, *377*, 331–345. [CrossRef]
16. Yao, Y.; Zhang, S.; Yang, S.; Gui, G. Learning Attention Representation with a Multi-Scale CNN for Gear Fault Diagnosis under Different Working Conditions. *Sensors* **2020**, *20*, 1233. [CrossRef]
17. Zhang, W.; Li, C.; Peng, G.; Chen, Y.; Zhang, Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mech. Syst. Signal Process.* **2018**, *100*, 439–453. [CrossRef]
18. Jia, F.; Lei, Y.; Lu, N.; Xing, S. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mech. Syst. Signal Process.* **2018**, *110*, 349–367. [CrossRef]
19. Wang, X.; Cui, L.; Wang, H.; Jiang, H. A generalized health indicator for performance degradation assessment of rolling element bearings based on graph spectrum reconstruction and spectrum characterization. *Measurement* **2021**, *176*, 109165. [CrossRef]
20. Mao, W.; Liu, Y.; Ding, L.; Li, Y. Imbalanced Fault Diagnosis of Rolling Bearing based on Generative Adversarial Network: A Comparative Study. *IEEE Access* **2019**, *7*, 9515–9530. [CrossRef]
21. Zhou, F.; Yang, S.; Fujita, H.; Chen, D.; Wen, C. Deep learning fault diagnosis method based on global optimization GAN for unbalanced data. *Knowl.-Based Syst.* **2020**, *187*, 104837. [CrossRef]
22. Yang, J.; Gao, T.; Jiang, S.; Li, S.; Tang, Q. Fault Diagnosis of Rotating Machinery Based on One-Dimensional Deep Residual Shrinkage Network with a Wide Convolution Layer. *Shock Vib.* **2020**, *2020*, 8880960. [CrossRef]
23. Yaqub, M.F.; Gondal, I.; Kamruzzaman, J. An Adaptive Self-Configuration Scheme for Severity Invariant Machine Fault Diagnosis. *IEEE Trans. Reliab.* **2013**, *62*, 160–170. [CrossRef]
24. Dong, Y.; Li, Y.; Zheng, H.; Wang, R.; Xu, M. A new dynamic model and transfer learning based intelligent fault diagnosis framework for rolling element bearings race faults: Solving the small sample problem. *ISA Trans.* **2022**, *121*, 327–348. [CrossRef] [PubMed]
25. Shao, S.; Wang, P.; Yan, R. Generative adversarial networks for data augmentation in machine fault diagnosis. *Comput. Ind.* **2019**, *106*, 85–93. [CrossRef]
26. Zhou, Q.; Li, Y.; Tian, Y.; Jiang, L. A novel method based on nonlinear auto-regression neural network and convolutional neural network for imbalanced fault diagnosis of rotating machinery. *Measurement* **2020**, *161*, 107880. [CrossRef]
27. Zhao, D.; Liu, S.; Gu, D.; Sun, X.; Wang, L.; Wei, Y.; Zhang, H. Enhanced data-driven fault diagnosis for machines with small and unbalanced data based on variational auto-encoder. *Meas. Sci. Technol.* **2020**, *31*, 035004. [CrossRef]
28. Zhu, T.; Lin, Y.; Liu, Y. Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recogn.* **2017**, *72*, 327–340. [CrossRef]
29. Han, H.; Wang, W.; Mao, B. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
30. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In *Pacific-Asia Conference on Advances in Knowledge Discovery & Data Mining*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 475–482.
31. He, H.; Bai, Y.; Garcia, E.A.; Li, S. *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning*; IEEE: Piscataway, NJ, USA, 2008; pp. 1322–1328.
32. Kıymık, M.K.; Güler, O.; Dizibüyük, A.; Akın, M. Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application. *Comput. Biol. Med.* **2005**, *35*, 603–616. [CrossRef]
33. Azuara, G.; Ruiz, M.; Barrera, E. Damage Localization in Composite Plates Using Wavelet Transform and 2-D Convolutional Neural Networks. *Sensors* **2021**, *21*, 5825. [CrossRef]
34. Chikkerur, S.; Cartwright, A.N.; Govindaraju, V. Fingerprint enhancement using STFT analysis. *Pattern Recogn.* **2007**, *40*, 198–211. [CrossRef]
35. Alexakos, C.T.; Karnavas, Y.L.; Drakaki, M.; Tziafettas, I.A. A Combined Short Time Fourier Transform and Image Classification Transformer Model for Rolling Element Bearings Fault Diagnosis in Electric Motors. *Mach. Learn. Know. Extr.* **2021**, *3*, 228–242. [CrossRef]
36. Gou, L.; Li, H.; Zheng, H.; Li, H.; Pei, X. Aeroengine Control System Sensor Fault Diagnosis Based on CWT and CNN. *Math. Probl. Eng.* **2020**, *2020*, 5357146. [CrossRef]
37. Yoo, Y.; Baek, J. A Novel Image Feature for the Remaining Useful Lifetime Prediction of Bearings Based on Continuous Wavelet Transform and Convolutional Neural Network. *Appl. Sci.* **2018**, *8*, 1102. [CrossRef]
38. Legendre, S.; Massicotte, D.; Goyette, J.; Bose, T.K. Wavelet-Transform-Based Method of Analysis for Lamb-Wave Ultrasonic NDE Signals. *IEEE Trans. Instrum. Meas.* **2000**, *49*, 524–530. [CrossRef]

39. Liang, P.; Deng, C.; Wu, J.; Yang, Z. Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network. *Measurement* **2020**, *159*, 107768. [CrossRef]
40. Zhou, K.; Sisman, B.; Li, H. Vaw-gan for disentanglement and recomposition of emotional elements in speech. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 415–422.
41. Kwon, H.; Kim, M.; Baek, J.; Chung, K. Voice Frequency Synthesis using VAW-GAN based Amplitude Scaling for Emotion Transformation. *KSII Trans. Internet Inf. Syst. (TIIS)* **2022**, *16*, 713–725.
42. Antoine, J.P.; Carrette, P.; Murenzi, R.; Piette, B. Image analysis with two-dimensional continuous wavelet transform. *Signal Process.* **1993**, *31*, 241–272. [CrossRef]
43. Vashishtha, G.; Kumar, R. Pelton Wheel Bucket Fault Diagnosis Using Improved Shannon Entropy and Expectation Maximization Principal Component Analysis. *J. Vib. Eng. Technol.* **2022**, *10*, 335–349. [CrossRef]
44. Jalayer, M.; Orsenigo, C.; Vercellis, C. Fault detection and diagnosis for rotating machinery: A model based on convolutional LSTM, Fast Fourier and continuous wavelet transforms. *Comput. Ind.* **2021**, *125*, 103378. [CrossRef]
45. Ferrante, M.; Brunone, B.; Meniconi, S. Wavelets for the Analysis of Transient Pressure Signals for Leak Detection. *J. Hydraul. Eng.* **2007**, *133*, 1274–1282. [CrossRef]
46. Halder, S.; Bhat, S.; Dora, B. Start-up transient analysis using CWT and ridges for broken rotor bar fault diagnosis. *Electr. Eng.* **2022**. [CrossRef]
47. Shao, H.; Xia, M.; Wan, J.; de Silva, C.W. Modified Stacked Autoencoder Using Adaptive Morlet Wavelet for Intelligent Fault Diagnosis of Rotating Machinery. *IEEE/ASME Trans. Mechatron.* **2022**, *27*, 24–33. [CrossRef]
48. Wang, H.; Li, S.; Song, L.; Cui, L. A novel convolutional neural network based fault recognition method via image fusion of multi-vibration-signals. *Comput. Ind.* **2019**, *105*, 182–190. [CrossRef]
49. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
50. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
51. Han, J.; Moraga, C. *The Influence of the Sigmoid Function Parameters on the Speed of Backpropagation Learning*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 195–201.
52. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
53. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. Acm.* **2017**, *60*, 84–90. [CrossRef]
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
55. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
56. Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
57. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information, Montreal, QC, Canada, 8–13 December 2014.
58. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
59. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Proc. Mag.* **2018**, *35*, 53–65. [CrossRef]
60. Li, W.; Zhong, X.; Shao, H.; Cai, B.; Yang, X. Multi-mode data augmentation and fault diagnosis of rotating machinery using modified ACGAN designed with new framework. *Adv. Eng. Inform.* **2022**, *52*, 101552. [CrossRef]
61. He, J.; Yang, S.; Papatheou, E.; Xiong, X.; Wan, H.; Gu, X. Investigation of a multi-sensor data fusion technique for the fault diagnosis of gearboxes. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2019**, *233*, 4764–4775. [CrossRef]

sensors

MDPI

Article

Transient Thermal Analysis Model of Damaged Bearing Considering Thermo-Solid Coupling Effect

Yali Sun [1], Chong Zhang [1], Xing Zhao [2], Xiaodong Liu [2], Chang Lu [1,3] and Jiyou Fei [2,*]

[1] College of Mechanical Engineering, Dalian Jiaotong University, Dalian 116028, China
[2] College of Locomotive and Rolling, Dalian Jiaotong University, Dalian 116028, China
[3] PLA Army Academy of Artillery and Air Defense, Shenyang 110000, China
* Correspondence: fjy@djtu.edu.cn

Abstract: As one of the important parameters of bearing operation, temperature is a key metric to diagnose the state of service of a bearing. However, there are still some shortcomings in the study of the temperature variation law for damaged bearings. In this paper, according to the structural characteristics of bearings, the influence law of thermal-solid coupling effect on bearing structure is considered, and a novel transient temperature analysis model of damaged bearings is established. First, a quasi-static analysis of the bearing is performed to obtain the variation laws of the key parameters of the bearing under thermal expansion. Then, the load variation law of the bearing under the condition of damage is discussed, and the heat generation and heat transfer of the damaged bearing during operation are studied. Based on the thermal grid method, a transient temperature analysis model of the damaged bearing is developed. Finally, the model is tested experimentally and the influence of the rotate speed and load on the bearing temperature variation is analyzed. The results show that the established model can effectively predict the temperature variation and thermal equilibrium temperature of damaged bearings.

Keywords: thermal-solid coupling; bearing fault diagnosis; heat generation and transfer; transient thermal model

Citation: Sun, Y.; Zhang, C.; Zhao, X.; Liu, X.; Lu, C.; Fei, J. Transient Thermal Analysis Model of Damaged Bearing Considering Thermo-Solid Coupling Effect. *Sensors* **2022**, *22*, 8171. https://doi.org/10.3390/s22218171

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 15 September 2022
Accepted: 21 October 2022
Published: 25 October 2022

1. Introduction

The rolling bearing is one of the most important components of high-speed trains, and plays a key role in in safety of transportation. However, the bearing easily produces defects, due to its hostile working environment [1]. Therefore, the detection of the state of bearings is quite important. However, the existing non-contact detection methods often ignore the influence of temperature on the important parameters of faulty bearings [2] (such as characteristic frequency). What should be noticed is that the faulty bearing is a system affected by the thermal-solid coupling effect. The structure size, contact load and deformation between components and lubricating oil characteristics will affect the heat generation and transfer of the bearing, and the temperature will promote the change of the features of oil, the size and performance of component structure and other parameters that affect bearing vibration characteristics. The bearing will achieve thermal equilibrium in the constant interaction and change. The contact load and deformation of the faulty bearings are quite complex, which makes the thermal analysis of such a system difficult, and leads to inaccurate basis of bearing detection and evaluation. Therefore, it is of great significance to understand the thermo-solid coupling effect of faulty bearings in operation and establish an accurate temperature prediction model for faulty rolling bearings.

Currently, research on bearing thermal analysis focuses on lubrication, heat generation and heat transfer. The bearings must be lubricated during operation, and the most important function of the lubricating oil is to provide a film of lubrication between the parts of the bearings. The thickness of the lubricant oil film depends on the surface roughness [3],

structure [4], load [5], speed [6] and temperature [7] of the bearing. Dowson and Higginson [8] proposed the classical oil film thickness calculation formula, which can calculate the oil film thickness under constant temperature based on elastic mechanics without considering the influence of temperature changes, and the calculated oil film thickness is slightly higher than the actual state [9]. Since then, several scholars have modified and refined this classical formula. For example, Forster et al. [10] and Echavarri Otero et al. [11] newly calculated the oil film thickness by adding empirical coefficients, and some scholars considered the non-Newtonian shear thinning effect when calculating the oil film thickness, such as Liu et al. [12] and Shirzadegan et al. [13]. In addition, some scholars have considered the influence of thermal effect, such as Wang et al. [14] and Echa-Varri Otero et al. [15].

The heat generated by the roll bearing comes mainly from friction heat production between the roll body and the racetrack, the roll body and the cage, and the inner loop guide surface and the cage. From this point of view, the heat generated by a bearing is mainly frictional, and therefore it will be affected by factors such as the bearing velocity, load and structural parameters. For the calculation method of heat generated, Palmgren [16] took the lead in proposing the global method of calculation, and obtained the empirical formula of bearing friction torque by experimental method, and multiplied the friction torque by the bearing speed to obtain the global heat production of the bearing. Since then, several scholars have developed improvements based on the global approach. For example, Stein and Tu [17] further analyzed the influence of preloading force based on the global method; Kim et al. [18] considered the influence of bearing load and speed on friction torque; in addition, other scholars have made improvements [19]. However, the global approach still has limitations, and the specific heating position of the bearings cannot be obtained in this way. Furthermore, some scholars proposed the local method to calculate the bearing temperature [20]. This method calculates the heat production at each contact position according to the mechanical and kinematic relations of the bearing, which can clarify the heat production and heat transfer rules of each part of the bearing, and the results are more accurate [21,22].

The heat distribution of rolling bearings mainly refers to temperature transfer, such as heat transfer and heat convection in bearings.

Bearing heat distribution mainly refers to temperature transfer, primarily heat transfer and heat convection. The mode of heat transfer and the final distribution of temperature will alter the lifetime of the bearing. The calculation methods for heat transfer and temperature distribution of bearings are mainly divided into finite element method [23] and thermal network method [24]. Compared with the finite element method, the thermal network method has advantages such as mesh density, settable nodes and easy setting of boundary conditions [25]. Some scholars have studied the number of nodes [26], the distribution form [27] and the boundary setting method [28] of the network, and further developed the transient thermal network model of the bearing system [29], which can forecast the bearing temperature more precisely than the original steady state model.

In summary, there have been some advances and achievements in existing thermal analysis models for bearings, but how to accurately judge the temperature rise and thermal equilibrium of bearings under damage conditions remains a difficult problem for bearing detection. In particular, most scholars disregard the variations of heat-affected bearing components in the construction of thermal analysis models for bearing, so that there is still some bias in the calculation process. Moreover, the effect of bearing damage on the temperature law is not taken into account. Nonetheless, in the actual operation of a bearing, there is an interplay between the bearing structure, lubrication properties, load, working conditions and other key factors, as well as the bearing temperature, under which the thermal equilibrium of the bearing is reached progressively. However, many factors drastically change after bearing damage, making the process of temperature rise and thermal balance of damaged bearings extremely complicated. Therefore, it is important to model the thermal analysis of the bearing by taking into account the thermal-solid coupling effects of the damage.

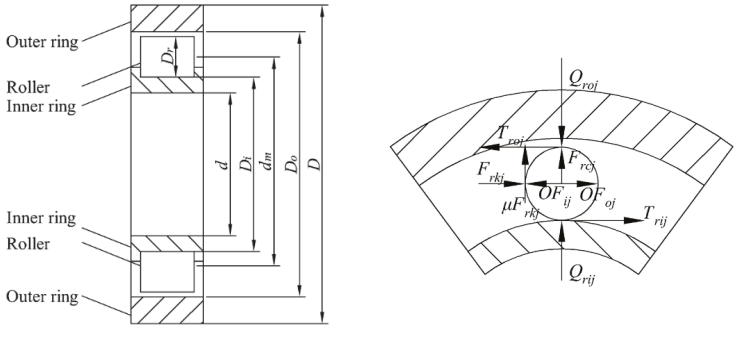
In this paper, an analytical model of temperature rises and thermal equilibrium temperature of injured bearings considering the influence of thermal-solid coupling effect is recommended. The interaction among structural parameters, contact load, heat generation, working conditions and the temperature of faulty rolling bearing was analyzed, and the predicting accuracy of the model was tested by verification experiments. This paper is structured as follows: the first section introduces the research status of thermal analysis model of rolling bearings in high-speed trains, and makes clear the difference of heat generation by contact load between good and faulty bearings. Section 2 describes the relationship between the structure parameters and force and motion of the roller, and proposes the load distribution calculation method of the faulty rolling bearing. Section 3 analyzes the calculation method of heat generation and transfer of faulty bearings during operation. Section 4 proposed the method of node division in thermal network and the calculation flow of transient temperature of the faulty rolling bearing. In Section 5, the corresponding verification experiments are carried out to prove the correctness of the temperature prediction model of the faulty bearing, and the influence of rotate speed and radial load on the temperature rise process is analyzed.

2. Dynamic Analysis of Damaged Bearings under the Influence of Thermal Expansion

2.1. Analysis of Bearing Motion and Dynamics

The accurate analysis of bearing mechanical characteristics is the premise of establishing a bearing thermal analysis model. The contact load, motion parameters and contact deformation among the internal components of high-speed rolling bearings have an important influence on the friction heat production in the process of operation. The accuracy of the calculation of these parameters will directly affect the accuracy of the thermal analysis model of bearings. Therefore, before establishing the bearing thermal analysis model, it is necessary to conduct accurate motion and force analysis on the rolling bearing first.

The schematic diagram of cylindrical rolling bearing is shown in Figure 1a. Assuming that the outer ring is fixed and there is no sliding phenomenon of the roller, the force analysis of a single roller is shown in Figure 1b. The rotational speed of the inner ring and roller are ω_i and ω_r. μ_{rij}, μ_{roj} and μ_{rkj} are friction coefficients between the jth roller and the inner ring, the outer ring and the cage, respectively. Q_{rij} and Q_{roj} are the contact loads between roller and inner ring and outer ring, respectively. F_{rkj} is the normal force between the roller and the cage, T_{rij} and T_{roj} are the friction force of the roller, OF_{ij} and OF_{oj} are the dynamic pressure of the roller from the lubricating oil.



(**a**) (**b**)

Figure 1. Schematic diagram of roller bearing: (**a**) schematic diagram of bearing structure; (**b**) the forces of a roller.

According to the motion relationship shown in the figure, the relative sliding velocities (ΔV_{rij}, ΔV_{roj}) and average velocities (V_{rij}, V_{roj}) of the jth roller with the inner and outer rings can be obtained, which can be expressed by the following formula:

$$\begin{cases} \Delta V_{rij} = \frac{1}{2} d_m \left[\left(1 - \frac{D_r}{d_m} \right) \omega_i - \frac{D_r}{d_m} \omega_r \right] \\ \Delta V_{roj} = -\frac{1}{2} D_r \omega_r \end{cases} \tag{1}$$

$$\begin{cases} V_{rij} = \frac{1}{4} d_m \left[\left(1 - \frac{D_r}{d_m} \right) \omega_i + \frac{D_r}{d_m} \omega_r \right] \\ V_{roj} = \frac{1}{4} D_r \omega_r \end{cases} \tag{2}$$

where d_m is the pitch diameter of the bearing and the D_r is the diameter of the roller.

Further, the force balance equations for the bearing under stable operation can be obtained, which can be expressed in the following formulae, respectively:

$$\begin{cases} OF_{ij} + T_{rij} - \left(OF_{oj} + T_{roj} \right) \pm F_{rkj} = 0 \\ Q_{rij} + F_{rcj} - Q_{roj} \pm \mu_{rkj} F_{rkj} = 0 \\ Q_r - \sum\limits_{j=1}^{Nr} Q_{ioj} \cos \theta_j = 0 \\ \sum\limits_{j=1}^{Nr} F_{rkj} = 0 \\ T_{rij} + T_{roj} - \mu_{rkj} F_{rkj} = 0 \end{cases} \tag{3}$$

where Q_r is the radial load of the rolling bearing, θ_j is the angular position of the roller, Nr is the number of the rollers, and F_{rcj} is the centrifugal force of the roller, which can be expressed by the following formula:

$$F_{rcj} = \frac{1}{2} m_r d_m \omega_r^2 \tag{4}$$

where m_r is the quality of the roller.

By taking into account the thermal effect and shearing-thinning, the contact loads between roller and the inner and outer ring can be solved by the following formula:

$$\begin{cases} Q_{rij} = K_{rij} \left(\delta_{rij} + 0.13 h_{rij} \right)^{10/9} \\ Q_{roj} = K_{roj} \left(\delta_{roj} + 0.13 h_{roj} \right)^{10/9} \end{cases} \tag{5}$$

where K_{rij} and K_{roj} are the stiffness coefficients of the inner and outer rings, respectively. δ_{rij} and δ_{roj} are the deformation degrees of contact. h_{rij} and h_{roj} are the minimum oil film thickness between the roller and the inner and outer ring, which can be obtained by the following formula:

$$\begin{cases} h_{rij} = 2.65 \frac{\varphi_t}{\varphi_{nc}} \left(\frac{\alpha_p^{0.54} \left(\eta_0 \Delta V_{rij} \right)^{0.7} R_{rij}^{0.43} L_r^{0.13}}{E^{0.03} Q_{rij}^{0.13}} \right) \\ h_{roj} = 2.65 \frac{\varphi_t}{\varphi_{nc}} \left(\frac{\alpha_p^{0.54} \left(\eta_0 \Delta V_{roj} \right)^{0.7} R_{roj}^{0.43} L_r^{0.13}}{E^{0.03} Q_{roj}^{0.13}} \right) \end{cases} \tag{6}$$

where η_0 and α_p are the dynamic viscosity and viscosity-pressure coefficients of oil, which are affected by the temperature. R_{rij} and R_{roj} are the equivalent curvature radius of the contact between roller and inner and outer rings, L_r is the length of roller, φ_t is the oil film thermal correction coefficient and φ_{nc} is the non-Newtonian fluid correction coefficient, which can be obtained according to the literature [30].

The total deformation of the rolling bearing can be expressed by the following formula:

$$\delta_{rj} = \left(\delta_{\max} + 0.5 H_{ol} - \left(h_{ri1} + h_{ro1} \right) \right) \cos \theta_j + \left(h_{rij} + h_{roj} - 0.5 H_{ol} \right) \tag{7}$$

where H_{ol} is the bearing clearance.

The bearing clearance H_{ol} in the thermal expansion state can be calculated as follows.

$$H_{ol} = H_0 + \Delta_r - \left(\Delta D_i + \Delta D_o + 2 \Delta D_r \right) \tag{8}$$

where H_o is the initial clearance of the bearing, Δ_r is the increment of the clearance after thermal expansion, and ΔD_i, ΔD_o and ΔD_r are the diameter changes of the rollers, inner and outer rings, caused by the change of the bearing temperature, which can be solved by the following formula:

$$\begin{cases} \Delta D_i = \dfrac{2(I_i+\Delta I_i)\left(\frac{D_i}{d}\right)}{\left[\left[\dfrac{\left(\frac{D_i}{d}\right)^2+1}{\left(\frac{D_i}{d}\right)^2-1}+v_i\right]+\frac{E_i}{E_a}\left[\dfrac{\left[\left(\frac{d}{d'}\right)^2+1\right]}{\left[\left(\frac{d}{d'}\right)^2-1\right]}-v_a\right]\right]\left[\left(\frac{D_i}{d}\right)^2-1\right]} \\ \Delta D_o = \dfrac{2(I_o+\Delta I_o)\left(\frac{D}{D_o}\right)}{\left[\left[\dfrac{\left(\frac{D}{D_o}\right)^2+1}{\left(\frac{D}{D_o}\right)^2-1}+v_b\right]+\frac{E_o}{E_b}\left[\dfrac{\left[\left(\frac{L_H}{D}\right)^2+1\right]}{\left[\left(\frac{L_H}{D}\right)^2-1\right]}-v_o\right]\right]\left[\left(\frac{D}{D_o}\right)^2-1\right]} \\ \Delta D_r = \alpha D_r(T_r - T_c) \\ \Delta_r = \alpha D_o(T_o - T_c) - \alpha D_i(T_i - T_c) \\ \Delta I_i = \alpha d(T_a - T_c) - \alpha d(T_i - T_c) \\ \Delta I_o = \alpha D(T_o - T_c) - \alpha D(T_b - T_c) \end{cases} \tag{9}$$

where D_i is the inner raceway diameter; D_o is the outer raceway diameter; D is the outside diameter; d is the bore diameter; d' is the shaft inner diameter; L_H is the housing outer diameter; α is the linear expansion coefficient; E and ν are elastic modulus and Poisson ratio; I and ΔI are interference and increments; T_c is the initial temperature; and T_r, T_i, T_o and T_b are, respectively, the temperatures of the roller, inner ring, outer ring, shaft and bearing housing.

Through the above analysis, the variation law of contact load and velocity of the roller and the relationship between the structural parameters and temperature can be obtained. However, the load distribution form of faulty bearings is obviously different from that of good bearings. Therefore, in order to more accurately calculate the temperature variation law of the faulty bearing during operation, it is necessary to further analyze the load distribution form of the faulty bearing.

2.2. The Analysis of Contact Load Distribution of Faulty Bearings

Take the outer ring as an example. Figure 2 shows the schematic diagram of faulty bearing with a defect on the outer ring. Assuming that the rolling bearing is rigidly supported, the defect depth is h, the defect width is 2b, θ_j is the angular position of the jth roller, $\Delta\theta_j$ is the angle corresponding to the defect width, θ_f is the angular position of the defect center and d_ψ is the actual depth of the roller into the defect.

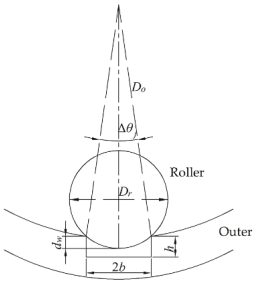


Figure 2. The interaction between the race and the defect.

Since there is a defect on the outer ring, the contact deformation between the roller and the outer ring should be analyzed in two parts [31]. The contact deformation between roller and good position of the outer ring has been analyzed by a large number of literatures, so it will not be discussed in this paper. The other case is that the roller is in contact with the defect. In the process of calculating load distribution in this situation, it is necessary

to estimate whether the *j*th rolling body is in the range of defect, or not. That is, the roller needs to meet the following formula:

$$-\frac{\Delta\theta_j}{2} < \mathrm{mod}(\frac{\theta_j}{2\pi}) - \theta_f < \frac{\Delta\theta_j}{2} \tag{10}$$

$$\Delta\theta_j = 2\arcsin\frac{b}{0.5D_o} \tag{11}$$

In this case, the contact deformation between the roller and the outer ring is:

$$\delta_{rj} = \delta_r \cos(\theta_j) - \frac{1}{2}H_{ol} - d_\psi \tag{12}$$

$$d_\psi = 0.5D_r - \sqrt{(0.5D_r)^2 - b^2} \tag{13}$$

According to Stribeck's theory, the relationship between the contact deformation and the maximum contact deformation $\delta_{r\max}$ can be obtained:

$$\delta_{rj} = \delta_{r\max}\left[1 - \frac{1}{2\varepsilon}(1 - \cos(\theta_j))\right] \tag{14}$$

where ε is the distribution coefficient, which is related to the bearing clearance H_{ol} and the total displacement δ_r, and can be obtained according to the method proposed in reference [32].

According to the load-deformation relationship between the rolling body and the outer ring, the relationship among contact load, maximum contact load, contact deformation and maximum contact deformation can be obtained:

$$Q_{rj} = Q_{\max}(\delta_{rj}/\delta_{r\max})^{10/9} = Q_{\max}\left[1 - \frac{1}{2\varepsilon}(1 - \cos(\theta_j))\right]^{10/9} \tag{15}$$

where the maximum contact load can be expressed as:

$$Q_{\max} = K_{rj}(\delta_r - 0.5H_{ol})^{10/9} \tag{16}$$

Since the radial load of the bearing is the sum of the loads of each roller, the following formula can be obtained:

$$Q_r = Q_{\max}\sum_{\theta_j=0}^{\theta_j=2\pi}\left[1 - \frac{1}{2\varepsilon}(1 - \cos(\theta_j))\right]^{10/9}\cos(\theta_j) \tag{17}$$

According to the above analysis, the load distribution and deformation of each rolling body of the faulty bearing in contact with the inner and outer ring can be obtained, which provides a theoretical basis for the analysis of the heat generation rate of each key component of the bearing.

3. Heat Generation and Transfer at Key Positions of Bearing

The heat generated by the bearing mainly comes from the friction between the surfaces of its parts. It includes the friction between the roller and the surface of the inner and outer ring, the friction between the roller and the cage and the friction between the guiding surface of the inner ring and the cage. Compared with the heat generated by other forms of friction, the heat generated by the friction between parts and oil is small and can be negligible, so only the friction heat generated by sliding friction is considered in this section. Normally, the heat generated by bearings is transmitted through heat conduction, heat radiation and heat convection. However, the bearing is installed in a relatively narrow

space, so less heat is transferred in the form of thermal radiation, which is not considered in this section.

3.1. Heat Generation in Bearing System

(1) Heat generation between roller and raceway.

The friction heat production between the roller and the inner and outer ring face can be expressed by the following formula:

$$Q_1 = \mu Q_{ri} \Delta V_{ri} + \mu Q_{ro} \Delta V_{ro} \tag{18}$$

In the above equation, the contact load and relative sliding velocity have been described above, μ is the friction coefficient, which is affected by temperature and is related to oil properties, and its calculation formula is [33]:

$$\mu = \left[\frac{4}{\pi p} + \frac{n_{oil}\alpha_p}{3} \left(e^{0.707 n_{oil}\alpha_p p_0} + 1.866 e^{0.259 n_{oil}\alpha_p p_0} + 0.134 e^{0.966 n_{oil}\alpha_p p_0} \right) \right] \left(\eta_0 \frac{\Delta V}{h_c} \right) G_r^{1-n} \tag{19}$$

where n_{oil} is the power-law exponent of oil, G_r is the shear modulus, h_c is 1.333 times of the minimum oil film thickness, which can be calculated by Equation (8).

(2) Heat generation between roller and cage.

The heat production between the roller and the cage is related to the friction force, and the friction force is limited by the normal force, friction medium, rotation speed and geometry parameters. Therefore, the heat production between a roller and the cage can be expressed by the following formula:

$$Q_2 = \frac{1}{2} D_{rj} \omega_{rj} F_{rkj} \mu \tag{20}$$

(3) Heat generation between the guide surface of the inner ring and cage.

There is sliding friction between the guide surface of the inner ring and the cage. However, due to the lubricating oil effect between the two parts, the heat generation is not very high, and can be expressed by the following formula:

$$\begin{cases} Q_k = \frac{1}{2} d_m F_{fk} \omega_r \\ F_{fk} = \frac{\eta_0 \pi w_k d_k^2 \omega_r}{d_{ik} - d_k} \end{cases} \tag{21}$$

where F_{fk} is viscous friction of the oil, w_k is the thickness of the cage and d_k and d_{ik} are the diameters of the guide surface of the cage and inner ring.

The heat generation of the key component in rolling bearing system can be calculated by the Equations (18), (20) and (21). The total heat generation can be calculated by $N_r Q_1 + N_r Q_2 + Q_k$.

3.2. Heat Transfer in Bearing System

The heat generated by bearings can usually be transferred by heat conduction and heat convection. Heat is transmitted through the contact between the roller, the inner ring, the outer ring and the cage, and the heat convection mainly occurs between inner ring, outer ring, roller and lubricating oil. The heat transfer relationship between various parts in the bearing is shown in Figure 3, and R is the thermal resistance. To determine whether there is heat transfer inside the bearing, it should be judged according to the value of thermal resistance of each bearing component.

3.2.1. The Conduction Thermal Resistance

According to the structural characteristics of bearing parts, the direction of heat conduction can be specifically divided into two kinds: one is conducted along the axial direction, and the other is conducted in the radial direction.

(1) Radial conduction thermal resistance.

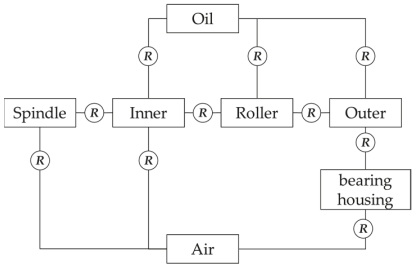


Figure 3. Relation diagram of heat transfer between the parts in bearing.

For the cylindrical parts such as roller and spindle, the conduction thermal resistance can be calculated by the following formula:

$$\theta = 1/\pi K_D L_r \tag{22}$$

where K_D is the thermal conductivity of the material and L_r is the effective length of the rolling body or shaft.

For the circular ring parts such as outer ring and inner ring, the conduction thermal resistance is:

$$\theta = \ln(r_i/r_o)/2\pi K_D L_{ring} \tag{23}$$

where r_i and r_o is the diameters of the simplified outer and inner rings, L_{ring} is the width of the simplified ring.

(2) Axial conduction thermal resistance.

For the calculation of the heat conduction resistance in the axial direction, only the shaft, bearing seat and rollers are considered, and the formula for the calculation is as follows:

$$\theta = L_{cc}/K_D A \tag{24}$$

where L_{cc} is the characteristic length of parts, A is the cross-sectional area.

(3) Contact conduction resistance between parts of the rolling bearing.

The most common contact form in bearings is the contact between the roller and the raceway. Based on the Hertz contact theory, the contact conduction resistance of this part can be expressed by the following formula:

$$\theta = \frac{1}{\pi}\left(\frac{a}{b}\right)\frac{1}{k_D a\sqrt{V_c C_p a\rho/k_D}} \tag{25}$$

where a is the major axis of the contact area, b is the minor axis of the contact area, V_c is the characteristic velocity and C_p is the specific heat capacity of the material.

The contact conduction resistance between inner ring and shaft, outer ring and bearing housing can be obtained according to the literature [23].

3.2.2. The Thermal Convection Resistance

In addition to thermal conduction, there are also thermal convection phenomena in the rolling bearing, including free and forced convection. For bearings, this process belongs to forced convection. Convective thermal resistance is related to Nusselt number, and its calculation formula is as follows:

$$\theta_w = L_{cc}/AK_w Nu \tag{26}$$

where K_w is the thermal conductivity of lubricating oil and N_u is the Nusselt number, which represents the strength of convective heat transfer.

Limited by different operating conditions and thermal convection conditions, the Nusselt numbers are different for different parts, and can be obtained according to the literature [26,33,34].

According to the above analysis of conductive and convective thermal resistance, the corresponding thermal resistance and calculation method of different parts of the faulty bearing system can be obtained by combining Figure 3.

4. Node Division and Calculation Process of the Model

4.1. Network Note Division

Based on the above analysis, according to the structural characteristics of double-row cylindrical roller bearing 130JRF05, the bearing structure is simplified and divided into 11 nodes on its structure, as shown in Figure 4. The node numbers and more details are as follows in Table 1. Due to the limitation of space, the calculation method of the transient temperature of each node can be found by referring to the literature [29].

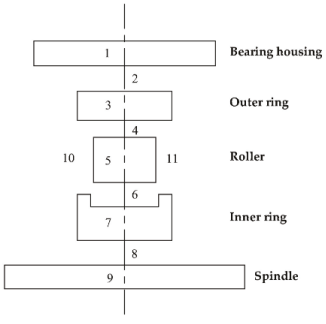


Figure 4. Node division of the rolling bearing system.

Table 1. The description of nodes.

Nobel Number	Description	Nobel Number	Description
1	Bearing housing	7	Inner ring
2	Contact of bearing housing and outer ring	8	Contact of inner ring and spindle
3	Outer ring	9	Spindle
4	Contact of outer ring and roller	10	Oil in bearing box
5	Roller	11	Air
6	Contact of roller and inner ring		

4.2. Transient Temperature Calculation Process of the Faulty Bearing

During the operation process of the rolling bearing, the final temperature distribution of the bearing system will be affected by the structure size, running speed, heat production, heat transfer and lubrication effect. In the process of increasing the temperature of the bearing, the thermal system of the bearing is a constantly changing and interacting system. For example, the structural size, working conditions and oil characteristics and other factors will affect the heat generation, and the change of temperature will in turn affect the bearing structure and oil parameters, until the bearing system reaches thermal balance. Therefore, the bearing thermal system is a thermo-solid coupling system. The transient temperature calculation flow of the faulty bearing is shown in Figure 5.

Firstly, the ambient temperature is set as the initial temperature for calculation, and the calculation time interval and bearing structural parameters are set. Secondly, the bearing

force is analyzed according to the working conditions, and then the bearing force and deformation under the initial conditions are obtained. Finally, the temperature variation law of bearing is obtained by combining the formula proposed in this paper. In the calculation process, the heat generation and transfer in first step are calculated, and then the obtained calculation results are used as the initial conditions for the next calculation step, and the iterative calculation is carried out continuously. The variable parameters affected by temperature are modified according to the results at the same time.

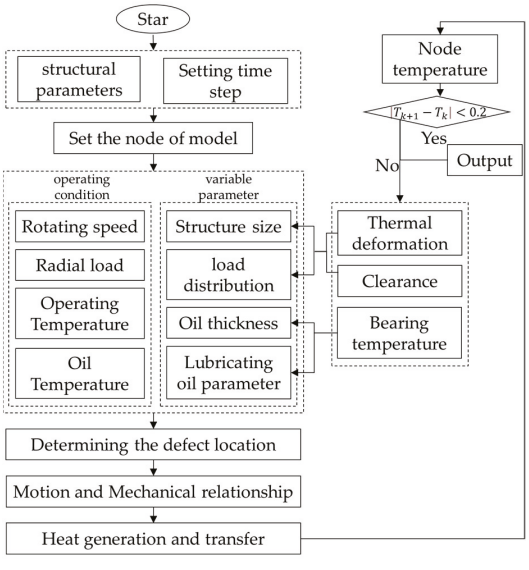


Figure 5. Transient temperature calculation flow of faulty bearing system.

5. Model Validation and Analysis

The proposed model is verified by a rolling bearing temperature measurement experiment platform. The basic parameters of oil are shown in Table 2, and the parameters that change under the influence of temperature are shown in Table 3. The bearings used are cylindrical roller bearings for high-speed trains produced by NSK Company, the model is N205, and its basic structural parameters are shown in Table 4.

Table 2. Basic parameters of the lubricating oil.

Parameters	Value
exponent sign (n_{oil})	0.43
shear modulus (G_r, Pa)	5.31×10^4
thermal conductivity (K_w, W/m °C^{-1})	0.132
density (ρ, kg/m^3)	920

The established experimental platform is shown in Figure 6. Through this platform, various states such as rotation and vibration of rotating machinery parts can be quickly simulated, and signals such as temperature, velocity and vibration can be collected. The schematic diagram of the platform structure and the sensor installation position are shown in Figure 7. The experimental platform is composed of an AC variable speed motor, a manual load adjusting device, an electric magnetic powder brake and bearing box, etc. The speed can be adjusted within 0~5000 r/min, and the load can be adjusted within 0~3 Kn. The acquisition card model is NIcDAQ-9178 reconstituted acquisition card, manufactured by National Instruments, Austin, TX, USA. The sampling device includes a dynamic

signal acquisition module, a bidirectional digital input module, a thermocouple signal acquisition module and a four-channel synchronous bridge signal acquisition module. During the experiment, the temperature is collected by digital channel. The components of the signal acquisition system are connected with each other by sensor signal lines, and there is no crossover phenomenon between lines. The defect of the faulty bearing used in the experiment is located in the outer ring, with a depth of 0.5 mm and a width of 3 mm. In the calculation, the initial temperature is set as the temperature of the experimental environment. In the experimental process, the temperature rise process and thermal equilibrium temperature of good and damaged bearings were compared under the same load and different speeds and under the same speed and different loads.

Table 3. Lubricating oil parameters after affected by temperature.

Temperature (°C)	Kinematic Viscosity (Ns/m^2)	Pressure–Viscosity Coefficient (m^2/N)	Dynamic Viscosity (m^2/s)
0	6.10×10^{-1}	3.03×10^{-8}	5.10×10^{-1}
10	3.12×10^{-1}	2.65×10^{-8}	3.05×10^{-1}
20	2.05×10^{-1}	2.13×10^{-8}	1.85×10^{-1}
30	1.31×10^{-1}	1.83×10^{-8}	1.23×10^{-1}
40	0.93×10^{-1}	1.75×10^{-8}	0.81×10^{-1}
50	0.47×10^{-1}	1.63×10^{-8}	0.32×10^{-1}
60	0.33×10^{-1}	1.57×10^{-8}	0.21×10^{-1}
70	0.19×10^{-1}	1.45×10^{-8}	0.12×10^{-1}
80	0.12×10^{-1}	1.32×10^{-8}	0.08×10^{-1}
90	0.08×10^{-1}	1.22×10^{-8}	0.04×10^{-1}

Table 4. Basic parameters of the N205.

Parameters	Value
Outerdiameter (mm)	52
Inner diameter (mm)	25
Width (mm)	15
Roller diameter (mm)	7.02
Pitch diameter (mm)	38.5
Number of Rollers	13
Bearing clearance (mm)	0.1
Interference (mm)	0.01

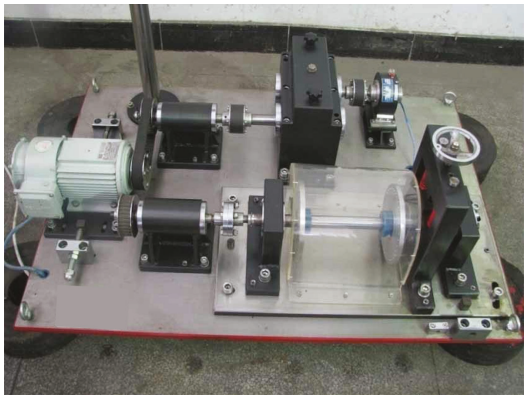


Figure 6. The bearing temperature acquisition and experiment platform.

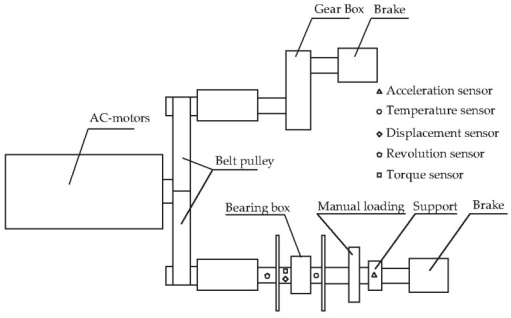
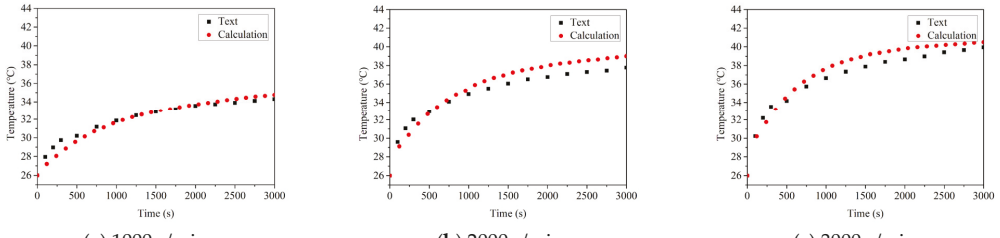


Figure 7. Schematic diagram of experimental platform structure.

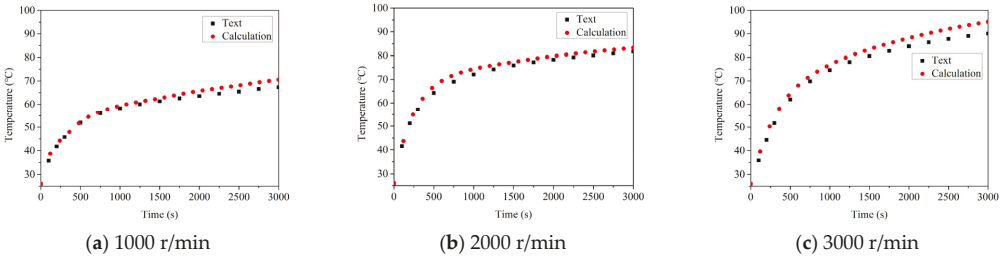
5.1. The Influence of the Rotational Velocity on Temperature

The temperature rises and thermal balance temperature of the good bearing and the defective bearing at the speed of 1000 r/min, 2000 r/min and 3000 r/min are compared, as shown in Figures 8 and 9, respectively. It can be found that the temperature rise process of good bearings is similar regardless of the speed of the bearings. Before the 2000 s, the temperature increased slowly and reached a stable state before 3000 s. However, the thermal balance temperature of the bearing will increase with the increase of the rotation speed. According to the previous analysis, the influence of speed on bearing temperature is very significant, and the increase of speed also leads to the increase of heat generation between parts of the rolling bearing, which leads to the higher temperature of bearing to reach thermal balance. As can be seen from the figure, when the speed reaches 3000 r/min, the bearing temperature increases by about 6 °C compared with that at 1000 r/min, which can also be observed from the experimental results.



(**a**) 1000 r/min (**b**) 2000 r/min (**c**) 3000 r/min

Figure 8. The influence of the speed on temperature change of complete bearing.



(**a**) 1000 r/min (**b**) 2000 r/min (**c**) 3000 r/min

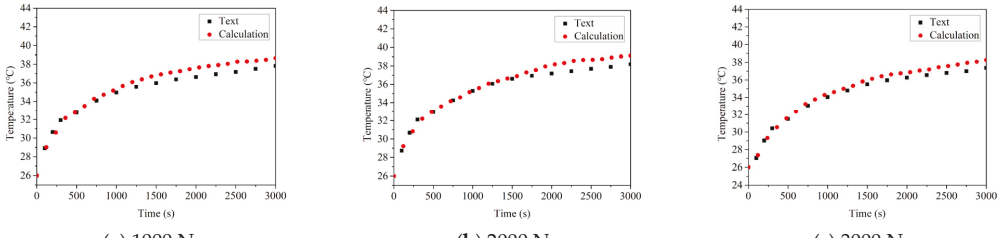
Figure 9. The influence of the speed on temperature change of damaged bearing.

Reviewing the temperature rise process of the damaged bearing, it can be found that the temperature is also rising in 1000 s of the bearing operation, but the climbing rate is obviously higher than that of the good bearing. After 1000 s, the temperature still rises, but

the rate of temperature rise decreases significantly, which is because the bearing structure is close to the limit of thermal expansion, leading to the gradual delay of heat production. At the same time, by comparing the thermal balance temperature of the two kinds of bearings at different speeds, it can be found that the thermal balance temperature of the faulty bearing is significantly higher than that of the good bearing. It is about 3~4 times that of the thermal equilibrium temperature of the good bearing, and the highest temperature reaches 95.06 °C, which is in good agreement with the experimental results. Through the error analysis of the results, the errors of good bearings at different speeds are 4.72%, 7.76% and 8.54%, respectively. The errors of faulty bearings at different speeds are 5.35%, 6.81% and 7.93%, respectively. The error is within the allowable range, which proves the correctness of the model.

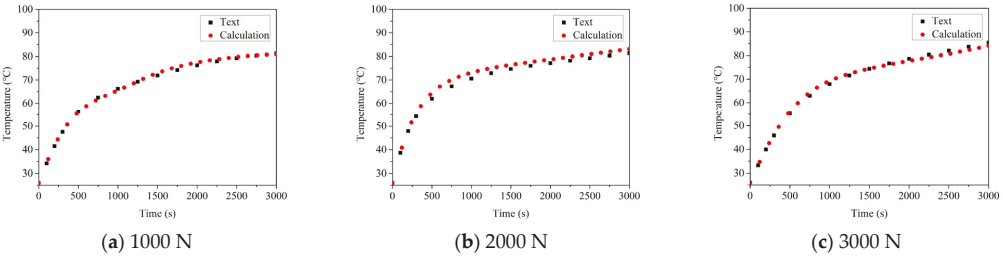
5.2. The Influence of the Load on Temperature

The temperature rises and thermal balance temperature of the good bearing and the defective bearing at the speed of 1000 N, 2000 N and 3000 N are compared, as shown in Figures 10 and 11, respectively. In the case of the good bearing, the temperature increases slowly with the running time of bearing, and finally reaches the thermal equilibrium state. Before reaching the stable state, the temperature rise rate does not change significantly. At the same time, it can be found that the temperature of the good bearing thermal equilibrium under different loads shows an upward trend, but the amplitude of temperature increase is small, which indicates that the radial load has no obvious influence on the temperature of the good bearing.



(**a**) 1000 N (**b**) 2000 N (**c**) 3000 N

Figure 10. The influence of the load on temperature change of complete bearing.



(**a**) 1000 N (**b**) 2000 N (**c**) 3000 N

Figure 11. The influence of the load on temperature change of damaged bearing.

However, another phenomenon can be found by observing the temperature rise of the faulty bearing. In the initial operation stage of the bearing, the temperature rises sharply, and the temperature rise rate can clearly see the boundary position. When the bearing structure is close to the limit size of thermal expansion, the temperature rise rate decreases. It can also be found that with the increase of load, the maximum temperature of the bearing gradually increases, and the amplitude of the lifting is significantly higher than that of the good bearing. However, compared with the influence of bearing speed on temperature of thermal equilibrium state, the influence of load is not very obvious. This

should be attributed to the fact that when the bearing has defects, with the increase of the radial load of the bearing, the contact load between the roller and the defect position will also be increased, which will result in the increase of heat generation between the roller and the outer ring, and then the temperature of the bearing will be increased. The same phenomenon can be also observed in the data obtained from the experiment. Based on the error analysis of the calculation results and experimental results, it can be found that the errors are 4.63%, 3.47% and 3.36%, respectively, under the condition of bearing without damage, and 2.87%, 3.12% and 2.66%, respectively, under the condition of bearing with defects. The error is also within the allowable range.

According to the experimental results and analysis above, the correctness of the established model is proved. The established bearing thermal analysis model, which, considering the influence of thermal-solid coupling, can effectively obtain the temperature variation law and thermal equilibrium temperature of damaged bearings under different operating conditions. It can also be found that when there is a fault in the bearing, the bearing speed has a more obvious influence on the bearing temperature compared with the radial load.

6. Conclusions

In this paper, a novel thermal analysis model of faulty bearing system is established, which considers the effects of thermal-solid coupling on the temperature variation. The influence of rotation speed and radial load on temperature variation of faulty bearing system is analyzed. The correctness of the established model is verified by experiments, and the conclusions are as follows:

(1) The coupling relationship between bearing heat generation and bearing structure, lubrication oil and working conditions was studied, and the transient temperature calculation model of faulty bearing was established with the method of thermal network. The heat generation rate and conduction resistance between friction surfaces of bearing were calculated, and the influence of thermo-solid coupling on temperature variation and thermal equilibrium temperature was obtained.

(2) Under the same working condition, the rate of heat generation and the thermal equilibrium temperature of the defective bearing are higher than those of the good bearing, because the defect changes the distribution of the contact load of the roller.

(3) Compared with the radial load, the rotate speed has a greater impact on the heat generation and thermal equilibrium temperature of the bearing, which is more obvious in the faulty rolling bearing.

The proposed model can effectively obtain the temperature of faulty bearings. The influence of temperature on vibration signal of faulty bearing can be further analyzed and applied to fault detection of bearing, such as acoustic detection, temperature detection and so on. However, this model still has some limitations: the lubricating oil of the bearing is constantly running and takes heat away, which is not considered in this paper. Therefore, the influence of flowing lubricating oil on the temperature of faulty bearings will be considered in the next study.

Author Contributions: Formal analysis, Y.S.; funding acquisition, X.Z.; software, Y.S., X.L. and C.L.; writing—original draft, Y.S.; writing—review and editing, C.Z. and J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Science Foundation for Young Scientists of China (Grant No. 62001079).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used to support the findings of this study are included within the article.

Acknowledgments: We would like to thank all the reviewers for providing valuable comments.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Wang, B.; Liu, Y.; Huai, W. Analysis of the Temperature Characteristics of High-speed Train Bearings Based on a Dynamics Model and Thermal Network Method. *Chin. J. Mech. Eng.* **2022**, *35*, 1–13. [CrossRef]
2. Jang, G.; Cho, S. Feature Space Transformation for Fault Diagnosis of Rotating Machinery under Different Working Conditions. *Sensors* **2021**, *21*, 1417. [CrossRef] [PubMed]
3. Kumar, R.; Ghosh, S.K.; Azam, M.S.; Khan, H. Numerical Simulation of Rough Thrust Pad Bearing Under Thin-Film Lubrication Using Variable Mesh Density. *Trans. Mech. Eng.* **2018**, *44*, 443–464. [CrossRef]
4. Wen, B.; Ren, H.; Dang, P.; Hao, X.; Han, Q. Measurement and calculation of oil film thickness in a ball bearing. *Ind. Lubr. Tribol.* **2018**, *70*, 1500–1508. [CrossRef]
5. Zhang, Z.; Wang, Y.; Lin, J.; Wang, D. Study on Factors Influencing Film Formation of Grease and Calculation Model for Grease Film Thickness. *Lubricants* **2022**, *10*, 123. [CrossRef]
6. Liu, Y.; Wang, B.; Zhang, B.; Yang, S. Establishment of Dynamic Model for Axle Box Bearing of High-Speed Trains Under Variable Speed Conditions. *Chin. J. Mech. Eng.* **2022**, *35*, 1–12. [CrossRef]
7. Nakahara, T.; Yagi, K. Influence of temperature distributions in EHL film on its thickness under high slip ratio conditions. *Tribol. Int.* **2007**, *40*, 632–637. [CrossRef]
8. Dowson, D.; Higginson, G.R. A Numerical Solution to the Elasto-Hydrodynamic Problem. *J. Mech. Eng. Sci.* **1959**, *1*, 6–15. [CrossRef]
9. Khonsari, M.M.; Kumar, P. EHL Circular Contact Film Thickness Correction Factor for Shear-Thinning Fluids. *J. Tribol.* **2008**, *130*, 4.
10. Forster, N.H.; Schrand, J.B.; Gupta, P.K. Viscoelastic Effects in MIL-L-7808-Type Lubricant, Part II: Experimental Data Correlations. *Tribol. Trans.* **2008**, *35*, 275–280. [CrossRef]
11. Echávarri Otero, J.; Lafont Morgado, P.; Chacón Tanarro, E.; De la Guerra Ochoa, E.; Diaz Lantada, A.; Munoz-Guijosa, I.; Munoz Sanz, I. Analytical model for predicting the friction coefficient in point contacts with thermal elastohydrodynamic lubrication. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2011**, *225*, 181–191. [CrossRef]
12. Liu, H.C.; Zhang, B.B.; Bader, N.; Bader, N.; Poll, G.; Venner, C.H. Influences of solid and lubricant thermal conductivity on traction in an EHL circular contact. *Tribol. Int.* **2020**, *146*, 106059. [CrossRef]
13. Shirzadegan, M.; Björling, M.; Almqvist, A.; Larsson, R. Low degree of freedom approach for predicting friction in elastohydrodynamically lubricated contacts. *Tribol. Int.* **2016**, *94*, 560–570. [CrossRef]
14. Wang, Y.; Cao, J.; Tong, Q.; An, G.; Liu, R.; Zhang, Y.; Yan, H. Study on the Thermal Performance and Temperature Distribution of Ball Bearings in the Traction Motor of a High-Speed EMU. *Appl. Sci.* **2020**, *10*, 4373. [CrossRef]
15. Echávarri Otero, J.; Guerra Ochoa, E.; Chacón Tanarro, E.; Díaz Lantada, A.; Munoz-Guijosa, J. Analytical model for predicting friction in line contacts. *Lubr. Sci.* **2016**, *28*, 189–205. [CrossRef]
16. Palmgren, A. *Ball and Roller Bearing Engineering*, 3rd ed.; SKF Industries Inc.: Philadelphia, PA, USA, 1959.
17. Stein, J.L.; Tu, J.F. A State-Space Model for Monitoring Thermally Induced Preload in Anti-Friction Spindle Bearings of High-Speed Machine Tools. *J. Dyn. Syst. Meas. Control.* **1994**, *116*, 43–53. [CrossRef]
18. Kim, K.; Lee, D.; Lee, S.; Lee, S.; Hwang, J. A numerical approach to determine the frictional torque and temperature of an angular contact ball bearing in a spindle system. *Int. J. Precis. Eng. Manuf.* **2015**, *16*, 135–142. [CrossRef]
19. Gao, P.; Tang, W.; Cui, Y.; Wang, Y.; Mo, G.; Yin, J. Theoretical and Experimental Investigation on Thermal Characteristics of Railway Double-Row Tapered Roller Bearing. *Energies* **2022**, *15*, 12. [CrossRef]
20. Zhou, X.; Zhang, H.; Hao, X.; Liao, X.; Han, Q. Investigation on thermal behavior and temperature distribution of bearing inner and outer rings. *Tribol. Int.* **2019**, *130*, 289–298. [CrossRef]
21. Mohan, L.; Gedong, J.; Xuesong, M.; Chit, M.; Jun, Y. Micro-Contact EHL Friction and Heat Generation Analysis of High Speed Ball Bearings. *Xi'an Jiaotong Univ.* **2016**, *50*, 81–88.
22. Ma, F.; Li, Z.; Wu, B.; An, Q. An accurate calculation method for heat generation rate in grease-lubricated spherical roller bearings. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2016**, *230*, 472–480. [CrossRef]
23. Xu, M.; Jiang, S.; Cai, Y. An improved thermal model for machine tool bearings. *Int. J. Mach. Tools Manuf.* **2006**, *47*, 53–62.
24. Ai, S.Y. Study on Temperature Distribution and Lubrication Performance of Rolling Bearings. Ph.D. Thesis, Beijing Institute of Technology, Beijing, China, 2015.
25. Li, M.; Lu, F.; Bai, X.; Zhu, W.; Zhu, R. Heat-flow coupling variation in double-row tapered-roller bearings during the loss of lubrication process. *Proc. Inst. Mech. Engineers. Part C J. Mech. Eng. Sci.* **2022**, *236*, 7500–7510. [CrossRef]
26. Zheng, D.; Chen, W. Thermal performances on angular contact ball bearing of high-speed spindle considering structural constraints under oil-air lubrication. *Tribol. Int.* **2017**, *109*, 593–601. [CrossRef]
27. Fangbo, M.; Zhengmei, L.; Shengchang, Q.; Baojie, W.; Qi, A. Transient thermal analysis of grease-lubricated spherical roller bearings. *Tribol. Int.* **2016**, *93*, 115–123.
28. Belmiloud, D.; Lachi, M.; Pron, H.; Bolaers, F.; Dron, J.; Chiementin, X.; Laggoun, A. Thermo-dynamical modelisation of the degradation of a ball bearing in variables use conditions. *Mech. Ind.* **2020**, *21*, 608–622. [CrossRef]

29. Liu, Y.; Wang, B.; Yang, S.; Liao, Y.; Guo, T. Characteristic analysis of mechanical thermal coupling model for bearing rotor system of high-speed train. *Appl. Math. Mech.* **2022**, *43*, 1381–1398. [CrossRef]
30. Ai, S.; Wang, W.; Wang, Y.; Zhao, Z. Temperature rise of double-row tapered roller bearings analyzed with the thermal network method. *Tribol. Int.* **2015**, *87*, 11–22. [CrossRef]
31. Liu, Y.; Wang, B.; Yang, S. Nonlinear Dynamic Behaviors Analysis of the Bearing Rotor System with Outer Ring Faults in the High-speed Train. *Chin. J. Mech. Eng.* **2018**, *54*, 17–25. [CrossRef]
32. Xu, H.; Xinxin, G.; Xianwen, Z.; Xin, L.; QingKa, H. Distribution characteristics of stress and displacement of rings of cylindrical roller bearing. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2019**, *233*, 151–160.
33. Pouly, F.; Changenet, C.; Ville, F.; Velex, P.; Damiens, B. Investigations on the power losses and thermal behaviour of rolling element bearings. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2010**, *224*, 925–933. [CrossRef]
34. Lee, J.; Kim, D.; Lee, C. A study on the thermal characteristics and experiments of High-Speed spindle for machine tools. *Int. J. Precis. Eng. Manuf.* **2015**, *16*, 293–299. [CrossRef]

sensors

MDPI

Article

Multiple Sensor Fault Detection Using Index-Based Method

Daijiry Narzary and Kalyana Chakravarthy Veluvolu *

School of Electronics and Electrical Engineering, Kyungpook National University, Daegu 41566, Korea
* Correspondence: veluvolu@ee.knu.ac.kr

Abstract: The research on sensor fault detection has drawn much interest in recent years. Abrupt, incipient, and intermittent sensor faults can cause the complete blackout of the system if left undetected. In this research, we examined the observer-based residual analysis via index-based approaches for fault detection of multiple sensors in a healthy drive. Seven main indices including the moving mean, average, root mean square, energy, variance, first-order derivative, second-order derivative, and auto-correlation-based index were employed and analyzed for sensor fault diagnosis. In addition, an auxiliary index was computed to differentiate a faulty sensor from a non-faulty one. These index-based methods were utilized for further analysis of sensor fault detection operating under a range of various loads, varying speeds, and fault severity levels. The simulation results on a permanent magnet synchronous motor (PMSM) are provided to demonstrate the pros and cons of various index-based methods for various fault detection scenarios.

Keywords: fault detection; fault detection index; residuals analysis; permanent magnet synchronous motor; multi-sensor faults

Citation: Narzary, D.; Veluvolu, K.C. Multiple Sensor Fault Detection Using Index-Based Method. *Sensors* **2022**, *22*, 7988. https://doi.org/10.3390/s22207988

Academic Editors: Dong Wang, Shilong Sun and Changqing Shen

Received: 8 September 2022
Accepted: 14 October 2022
Published: 19 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Sensors are frequently employed to gather data and signals, in particular in the monitoring of electrical devices and drives, the environment, and human health [1,2]. For instance, sensors are used in electrical motor drives to measure and detect changes in position, temperature, displacement, electrical current, as well as many other characteristics [3,4]. However, industrial sensors' applicability relies on the applications and conditions in which they are utilized. They are required to perform under challenging situations, such as severe and extreme environments with extremely low or high temperatures, vibrations, excessive humidity, etc. [5].

Any industrial drive's efficiency is entirely dependent on the output of the sensor's readings. An unexpected variation in the measured signal output, however, may be referred to as a sensor malfunction [6]. There are several causes of sensor faults, including poor manufacturing practices, long-term use wear and tear, and incorrect calibration. This frequently leads to physical divergence from the sensor body's design parameters, producing misleading and incorrect outputs [7]. Bias, drift, scaling, noise, and hard faults including signal loss are the main causes of sensor malfunction [8]. Different sensors, including voltage, current, temperature, pressure, and position sensors, are typically used in fault detection and diagnosis schemes [9]. The gathered sensor data reveal important details about the system's health, including whether it is functioning normally or not. Sensor fault diagnosis can be broadly divided into two main categories: hardware method and software method [10]. The hardware method uses multiple components and the same input signals, which are further utilized for comparison, and specific methods such as voting and limit test, etc., are utilized for fault detection. On the other hand, the software method is subdivided into model-based [11], signal-processing-based [12,13], and knowledge-based detection methods [14]. Any type of sensor fault can deteriorate the overall performance of an industrial drive by reducing its reliability. Therefore, it is necessary to investigate the sensor fault diagnosis of the drives, in order to ensure continuous drive operations [15,16]. Model-based methods [17–20] detect sensor faults by monitoring

the residual signal, which is the difference between the real process and the analytical redundancy under normal working conditions. They are considered as the most common techniques in industrial applications. The common residual generation methods include observer-based methods [21,22], parity space methods [23], and parameter identification methods for effectively detecting the sensor faults in satellite control systems and industrial motor drives. However, in signal processing methods for sensor fault detections, a faulty sensor signal of a motor is analyzed with signal processing techniques such as the fast Fourier transform (FFT) [24], wavelet transform (WT) [25], and Hilbert transform [26]. In [27], the STFT and WT methods were used for fault diagnosis such as demagnetization, rotor eccentricity faults, and sensor faults of the servo drive. Recent studies have used the signal processing fault diagnosis techniques focusing on the current, motor vibration, and voltage signals. In [28], the short-time-Fourier-transform (STFT)-based inverter fault detections were used for spectral analysis to detect open-circuit faults in a wind power converter. The knowledge-based method uses the summary of prior knowledge to describe the relationship between the fault and symptom. Interturn short-circuit faults in a drive were detected by using support vector machines (SVMs) and convolutional neural networks (CNNs) in [29]. In [30], bipolar transistor faults, single current sensor fault, and rotor position faults were diagnosed by the FDI algorithm designed by using the SVM technique. In [31], the demagnetization fault was identified using noise and torque information fusion technologies. Similarly, in [32], a Kalman-filter-based sensor fusion method was used to simultaneously measure the three-degree-of-freedom angular displacements and velocity of a ball-joint-like permanent magnet spherical motor.

Nevertheless, so far, the majority of the detection techniques rely only on data from one or more sensors. In the above methods, simultaneous or sequential faults in multiple sensors such as abrupt, incipient, and intermittent faults [33] were not discussed. Hence, in this work, three types of sensor faults such a abrupt, incipient, and intermittent faults are detected by the response of the indices generated from the fusion of speed, current, and voltage sensor residuals. Abrupt faults are modeled as a sudden step-like deviation in which the component value abruptly changes from its nominal value to an unknown faulty one. Incipient faults develop slowly, and intermittent faults usually manifest themselves intermittently in an unpredictable manner. Usually, abrupt faults and incipient faults have a persistent nature, while intermittent faults do not. In this paper, an intermittent fault [34,35] was considered to be periodic with a fixed value. Existing studies employed multiple sensors for the same sensor channel to reduce the noise and improve the fault detection accuracy by sensor fusion. Our proposed approach fundamentally differs in the way that we rely on a single sensor for one sensor channel. As a fault in one sensor channel affects the other sensor channels, we employed the index-based methods to analyze and identify the faulty and healthy channel.

In this work, finite time observers were employed for residual generation for analysis with various index-based methods. The drive was assumed to be healthy, and the issue of faults in multiple sensors was studied in the paper. Multiple sensors' fault detections based on indices were designed by using the moving root mean square index (MRI), moving-average-based index (MAI), moving-variance-based index (MVI), moving-energy-based index (MEI), first-order-derivative-based index (DBI_1), second-order-derivative-based index (DBI_2), and auto-correlation-based index (A_cI). An auxiliary index (AI) was also developed to select the accurate index values for faulty sensor detections. These index-based techniques provide quick and accurate fault detection. Cost effectiveness was also achieved by the extremely low computational burden of these index-based methods. For evaluation, the index-based fault detection methodology was tested on a permanent magnet synchronous motor (PMSM) with multiple sensors that were employed for speed, current, and voltage measurement. The simulation results are presented together with descriptions of the index-based detections for various defective and noisy settings. A comparative result is also presented to show the efficacy of the proposed method.

2. Index-Based Methods

To achieve accurate fault detection in multi-sensor faults, the following indices were considered for the analysis.

2.1. Moving-Average-Based Index

The moving average for a signal $p(t)$ can be calculated as follows:

$$MAI_i = \frac{1}{T_s} \sum_{n=i-T_s}^{i} p_i(t) \tag{1}$$

where MAI_i is the mean of the signal in the ith window. T_s is the number of samples in one cycle, and t is the time step of one sample. Each second moving average for each sample of a signal for the mean of a window of 1 s was calculated on the sample. The transients of the sensor residuals were analyzed by using the index-based methods. The MAI_i value remains constant during the PMSM motor's healthy sensor conditions, but it changes immediately after the fault occurs. Because of this, a threshold was used to compare the MAI_i and determine whether the index indicated the presence of faulty sensors. The value of the threshold for this index was considered as 0.5, by using Otsu's thresholding method.

2.2. Moving-RMS-Based Index

For abrupt and incipient sensor faults, the faults on one sensor affect the residuals of the other sensors. Hence, to detect the actual faulty sensor, the moving-RMS-based index is calculated as follows:

$$MRI_i = \sqrt{\frac{1}{T_s} \sum_{n=i-T_s}^{i} p_n^2} \tag{2}$$

where MRI_i is the root mean square of the signal $p(t)$ in the nth window. In this article, the root mean square was applied to the residuals of the stator currents, speed, and stator voltage values, with the number of samples in one cycle denoted as T_s and the time step of one sample as t. The object calculates the root mean square (RMS) of the windowed data at each iteration through the window. It can also be seen that the energy of the signal is directly proportional to the MRI values of the residuals, considering a constant window length. Like the behavior of the MAI, the MRI exhibits smooth fluctuations during the healthy sensor state, but it indicates a change during abrupt and incipient faults. The index for MRI, like the index MAI, was compared with a threshold designed using Otsu's thresholding. The threshold's considered value was 0.5.

2.3. Moving-Variance-Based Index

A variance-based index was used to separate the faulty sensor from the non-faulty ones by comparing it with the set threshold. Utilizing the formula of the moving average, MVI_i is considered as follows:

$$MVI_i = \frac{1}{T_s} \sum_{n=i-T_s}^{i} (p_n - \bar{p}_n)^2) \tag{3}$$

where p_n denotes each sample of the sensor residual and $\bar{p}_n$ is the average of the samples of the residuals in the specified window.

The moving variance calculates the variance of the signal around the mean in the given window. When a fault occurs, the abrupt changes cause large deviations, and this affects the variance of the signal.

2.4. Moving-Energy-Based Index

Abrupt and incipient faults are also detected by another index, called the moving-energy-based index. The index is calculated and then compared with a threshold.

This index can be calculated as follows:

$$MEI_i = \frac{1}{T_s} \sum_{n=i-T_s}^{i} p_n^2 \tag{4}$$

where MEI_i is the moving energy of the signal denoted by S_n. This index, like the MVI, exhibits the same behavior as the previous three indices. The comparison threshold was set at 0.2 and was created using Otsu's thresholding technique [36]

2.5. First-Order-Derivative-Based Index (DBI_1)

In this work, residual-based fault analysis was performed by designing a first-order-derivative-based index. It can be calculated as follows:

$$DBI_{1i} = \lim_{x_2 \to x_1} \frac{f(x_2) - f(x_1)}{x_2 - x_1} \tag{5}$$

The idea behind using a (DBI_1)-based index is to amplify the very slight changes in the faulty sensor values, as well as the noises present in the sensor residual transients. This index was compared to a threshold of 0.5 calculated using Otsu's thresholding method.

2.6. Second-Order-Derivative-Based Index (DBI_2)

A second-order-derivative-based index was also designed for the analysis of residuals for fault detections. It can be calculated as shown below:

$$DBI_{2i} = \lim_{x_2 \to x_1} \frac{\frac{d}{dx} f(x_2) - \frac{d}{dx} f(x_1)}{x_2 - x_1} \tag{6}$$

The DBI_2 index method, like the DBI_1 index method, analyzes the transients of noisy abrupt and incipient faults. The threshold was set to 0.5 and was created with Otsu's thresholding method.

2.7. Auto-Correlation Index

Another index utilized here for sensor fault detection was the auto-correlation index. This index was used for analyzing the residuals for intermittent faults. The mathematical expression for (A_cI) is shown as

$$A_cI = \frac{\sum_{i=k+1}^{n} (y_i - \bar{y})(y_{i-k} - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7}$$

where A_cI is the auto-correlation of the signal for time series of the signal y_i, and it lies between -1 and $+1$. $\bar{y}$ is the overall mean; n is the total number of samples; y_i is the value of the signal at sample i.

2.8. Auxiliary Index

In order to detect the sensor faults more accurately and preserve the reliability of the index-based methods, an auxiliary condition was considered by using an auxiliary index (AI). The mathematical representation of the auxiliary index is as follows:

$$AI = V_{indices}(f) > V_{indices}(nf) \tag{8}$$

where $V_{indices}$ are MRI, MAI, MVI, MEI, DBI_1, DBI_2, and A_cI, respectively; f indicates the faulty and nf indicates the non-faulty value of the indices.

Moreover, for the quantitative analysis of the proposed indices, the two following criteria, the accuracy *(Acc)* and dependability (*Dep*) of the indices, were calculated by using the formulae as follows:

$$Acc\% = \frac{\text{Number of correctly detected cases}}{\text{Total number of cases}} \tag{9}$$

$$Dep\% = \frac{\text{Total number of detected faults by the indices}}{\text{Total number of faults}} \tag{10}$$

3. Multi-Sensor Fault Diagnosis

In this paper, for the evaluation, we employed the proposed methodology for multi-sensor fault diagnosis in a (PMSM) motor. The dynamics of a PMSM can be modeled as follows:

$$\begin{cases} \frac{di_\alpha}{dt} = -\frac{R}{L}i_\alpha - \frac{1}{L}b_\alpha + \frac{1}{L}E_\alpha \\ \frac{di_\beta}{dt} = -\frac{R}{L}i_\beta - \frac{1}{L}b_\beta + \frac{1}{L}E_\beta \\ \frac{d\omega_e}{dt} = -\frac{P}{J}\phi_e(-\sin\theta_e i_\alpha + \cos\theta_e i_\beta) - \frac{F_e}{J}\omega_e - \frac{\Delta_e}{J} \\ \frac{d\theta_e}{dt} = \omega_e \end{cases} \tag{11}$$

where i_α and i_β are the stator currents, E_α and E_β are the stator voltages, and b_α and b_β are the back EMFs given as $b_\alpha = -K_e\omega_e \sin\theta_e$ and $b_\beta = K_e\omega_e \cos\theta_e$, respectively. In the above equations, R is the stator resistance, L is the synchronous inductance, P is the number of pole pairs, J is the moment of inertia, K_e is the back EMF constant, ϕ_e is the rotor flux, F_e is the viscous friction, Δ_e is the load torque, and θ_e and ω_e are the position and speed of the motor, respectively. The specifications of the motor parameters are defined in Table 1, and the functional block diagram of a PMSM is shown in Figure 1. In this work, abrupt, incipient, and intermittent faults were considered in the speed, current, and voltage sensors of a PMSM. Higher-order sliding mode (HOSM) observers were designed to generate the residuals of the speed and voltage sensors, respectively. However, a Luenberger observer was designed to generate the residuals of the current sensors. The main objective lied in the multi-sensor fault detection of a PMSM. The second objective was to validate the proposed method accordingly.

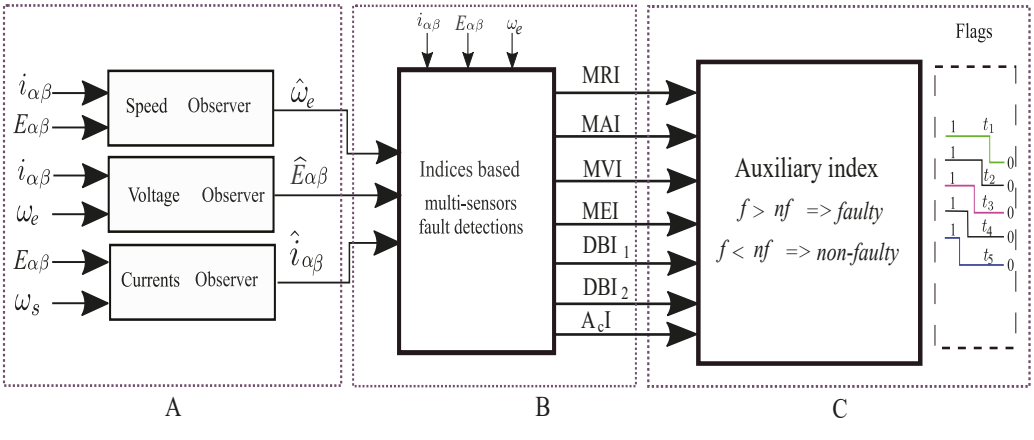


Figure 1. Functional block diagram for multi- sensor fault detection in a PMSM; (**A**): Finite time observer block; (**B**): Indices based detections using the residuals; (**C**): Fault Identification with the AI.

Table 1. Specifications of the PMSM.

Quantity	Symbol	Value
PMSM Rating	P_w	50 (kW)
Rating speed	ω_e	628 (rad/s)
Stator inductance	L	0.47 (mH)
Stator resistance	R	0.79 (Ω)
Magnetic flux linkage	K_e	0.2709 (Vs/rad)
Number of poles	P	4

3.1. Generation of Speed Sensor Residuals

In this section, stator voltages (E_α, E_β) and currents (i_α, i_β) are considered as known quantities and speed (ω_e) is considered as an unknown quantity.

$$\frac{d\hat{i}_{\alpha_1}}{dt} = -\frac{R}{L}\hat{i}_{\alpha_1} + \frac{1}{L}E_\alpha + \frac{1}{L}\lambda_1 \tag{12}$$

$$\frac{d\hat{i}_{\beta_1}}{dt} = -\frac{R}{L}\hat{i}_{\beta_1} + \frac{1}{L}E_\beta + \frac{1}{L}\lambda_2 \tag{13}$$

where λ_1 and λ_2 are the higher-order terms of STA and can be written as:

$$\begin{cases} \lambda_1(t) = -K_{s_1}\zeta_1(\alpha_s(t)) - K_{s_2}\int_0^t \zeta_2(\alpha_s(t))d\tau \\ \lambda_2(t) = -K_{s_1}\zeta_1(\beta_s(t)) - K_{s_2}\int_0^t \zeta_2(\beta_s(t))d\tau \end{cases} \tag{14}$$

where α_s and β_s are the selected sliding surfaces and

$$\zeta_1(\alpha_s(t)) = \alpha_s(t) + K_{s_3}\lceil \alpha_s \rfloor^{\frac{1}{2}} \tag{15}$$

$$\zeta_2(\alpha_s(t)) = \alpha_s(t) + \frac{K_{s_4}^2}{2} sign(\alpha_s(t)) + 1.5\lceil \alpha_s \rfloor^{\frac{1}{2}} \tag{16}$$

where $K_{s_1}, K_{s_2}, K_{s_3}$, and K_{s_4} are properly designed constant terms. Similarly, the terms $\zeta_1(\beta_s(t))$ and $\zeta_2(\beta_s(t))$ can be designed by replacing α_s with β_s. The estimation error dynamics can be defined as $\alpha_s(t) = \hat{i}_\alpha - i_\alpha$ and $\beta_s(t) = \hat{i}_\beta - i_\beta$ and can be computed similar to [37].

Using the estimated back EMF voltages, the speed of the PMSM can be computed as follows:

$$\hat{\omega}_e = \frac{1}{K_s}\sqrt{\hat{b}_\alpha^2 + \hat{b}_\beta^2} \tag{17}$$

where $\hat{b}_\alpha = K_{s_2}\int_0^t \zeta_2(\beta_s(t))d\tau$ and $\hat{b}_\beta = K_{s_2}\int_0^t \zeta_2(\beta_s(t))d\tau$, respectively. The speed sensor residuals can be computed as $\omega_{res} = \hat{\omega}_e - \omega_e$.

3.2. Generation of Voltage Sensor Residuals

In this section, the stator currents (i_α, i_β) and speed (ω_e) are treated as known quantities and voltages (E_α, E_β) are treated as unknown quantities. By using the STA-based HOSM observers, the stationary voltages are estimated in the α and β axes, respectively.

$$\begin{cases} \frac{d\hat{i}_{\alpha_2}}{dt} = -\frac{R}{L}\hat{i}_{\alpha_2} - \frac{1}{L}b_\alpha + \frac{1}{L}\lambda_3(t) \\ \frac{d\hat{i}_{\beta_2}}{dt} = -\frac{R}{L}\hat{i}_{\beta_2} - \frac{1}{L}b_\beta + \frac{1}{L}\lambda_4(t) \end{cases} \tag{18}$$

where $\lambda_3(t)$ and $\lambda_4(t)$ are the gains of the STA observer and can be defined as follows:

$$\begin{cases} \lambda_3(t) = -K_{v_1}\zeta_3(V_{1s}(t)) - K_{v_2}\int_0^t sign(V_{1s}(t))d\tau \\ \lambda_4(t) = -K_{v_1}\zeta_4(V_{2s}(t)) - K_{v_2}\int_0^t sign(V_{2s}(t))d\tau \end{cases} \tag{19}$$

with

$$\zeta_3(V_{1s}(t) = V_{1s}(t) + K_{v_3}\lceil V_{1s}\rceil^{\frac{1}{2}} \tag{20}$$

$$\zeta_4(V_{1s}(t) = V_{1s}(t) + \frac{K_{v_4}^2}{2} sign(V_{1s}(t)) + 1.5\lceil V_{1s}\rceil^{\frac{1}{2}} \tag{21}$$

where V_{1s} and V_{2s} are the sliding surfaces, respectively, and K_{v_1} and K_{v_2} are the STA gains. The estimation error dynamics from Equations (11) and (18) can be computed from [37].

The unknown voltages can be estimated as follows:

$$\begin{cases} \hat{E}_\alpha = -K_{v_2} \int_0^t \zeta_2(V_{1s}(t)d\tau \\ \hat{E}_\beta = -K_{v_2} \int_0^t \zeta_2(V_{2s}(t)d\tau \end{cases} \tag{22}$$

Hence, the voltage sensor residuals can be computed as $E_{\alpha\beta}, res$=$\hat{E}_{\alpha\beta} - E_{\alpha\beta}$.

3.3. Generation of Stator Current Sensor Residuals

In this section, the stator currents (i_α, i_β) are treated as unknown quantities and speed (ω_e) and voltages (E_α, E_β) are treated as known quantities. By using the Luenberger observer, the unknown stator currents are estimated in both the α and β axes. We utilize the PMSM model in the stationary reference frame as

$$\begin{cases} \frac{dx}{dt} = A_1 x(t) + B_1 u(t) \\ y(t) = C_1 x(t) \end{cases} \tag{23}$$

where x=$[i_\alpha, i_\beta, \omega_e, \theta_s]^T$ is the state vector. $u = [E_\alpha, E_\beta, T_l]^T$ and $y = [\omega_s, \theta_s]$ are the voltages and the input vector, respectively.

$$A_1 = \begin{bmatrix} -\frac{R}{L} & 0 & \frac{1}{L}P_k sin\theta_s & 0 \\ 0 & \frac{R}{L} & -\frac{1}{L}P_k cos\theta_s & 0 \\ -\frac{P}{J}\phi_m cos\theta_s & -\frac{F}{J} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} -\frac{1}{L} & 0 & \frac{1}{L} \\ 0 & \frac{1}{L} & 0 \\ 0 & 0 & -\frac{1}{J} \\ 0 & 0 & 0 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Th Luenberger observer can be designed as follows:

$$\frac{d\hat{x}}{dt} = A_1\hat{x} + B_1 u + L_t(y - C\hat{x}) \tag{24}$$

where $\hat{x} = [\hat{i}_\alpha, \hat{i}_\beta, \hat{\omega}_e, \hat{\theta}_s]^T$ is the state estimation vector and L_t is the observer gain matrix. The current sensor residuals can be computed as $i_{\alpha\beta,res} = \hat{i}_{\alpha\beta} - i_{\alpha\beta}$.

4. Results and Performance Evaluation

This section presents the simulation results to evaluate and demonstrate the effectiveness of the proposed index-based multi-sensor fault detections under different conditions. The specification of the parameters of the PMSM is mentioned in Table 1. The gain values of the HOSM observers for residual generations were selected as shown in [38]. The Luenberger observer (LO) gain matrix, L_t, can be selected from [39]. As shown in Figure 1, A represents the output of the finite time observers. The generated residuals as shown in Figure 1, B were further used for fault analysis. Hence, the indices MAI, MRI, MVI,

MEI, DBI_1, DBI_2, and A_cI were used for multiple sensors' fault detections. As shown in Figure 1, C an auxiliary index was used to differentiate the faulty sensor indices (f) from the non-faulty sensor indices (nf). The indices mentioned above can individually detect the sensor faults. Moreover, to improve the reliability and accuracy of the proposed method, the AI was used by collectively considering the indices and differentiating it based on the higher number of either faulty or non-faulty indices. The sensor faults in the PMSM motor can be classified as abrupt faults, incipient faults, and intermittent faults. In order to accurately detect the faults in the sensors, index-based analysis was performed by considering different types of sensor faults.

4.1. No-Fault Scenario

As shown in Figure 2, a speed reference of 1000 rpm was considered. The original and the estimated signals of the i_α and i_β currents, speed, and E_α, E_β voltage sensors are shown in Figure 2a(i), 2a(ii), 2a(iii), 2a(iv), and 2a(v), respectively. The corresponding residuals for the speed, stator voltages, and stator currents are shown in Figure 2b(i), 2b(ii), 2b(iii), 2b(iv), and 2b(v), respectively.

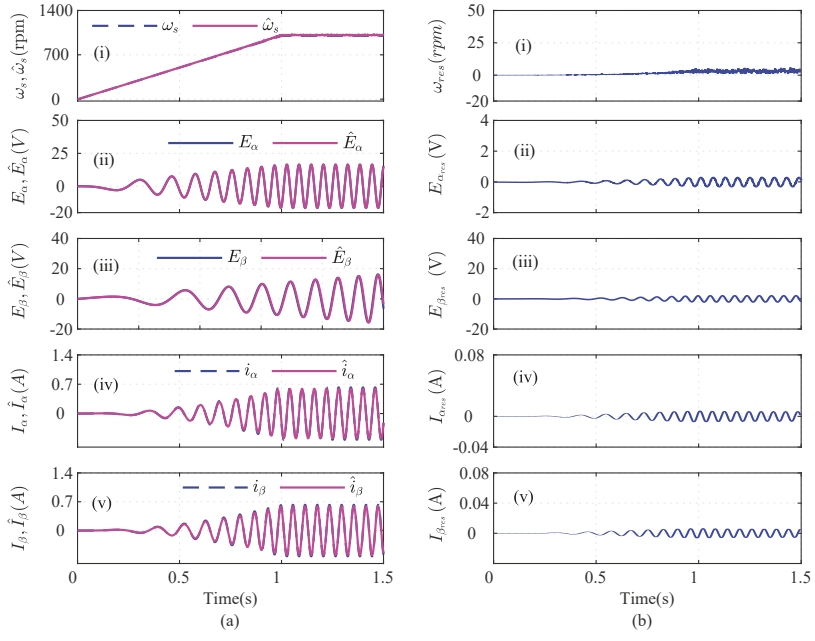


Figure 2. Illustration of current, speed, and voltage sensors' observers during no-fault scenario; (**a**) actual and estimated signals; (**b**) residuals.

4.2. Multi-Sensor Fault Scenario

In this section, multi-sensor faults are considered in the speed, current, and voltage sensors of a PMSM motor. In the first case, low-severity α, β abrupt current sensor faults with 15% load and low speed were introduced at t = 0.739 s and t = 1.52 s, as shown in Figure 3a(i) and Figure 3a(ii), respectively. The responses of the speed and voltage sensors are shown in Figure 3a(iii), 3a(iv) and 3a(v), respectively. The residuals of the α-, β-axis current, speed, and α-, β-axis voltage are shown in Figure 3b(i), 3b(ii), 3b(iii), 3b(iv), and 3b(v), respectively. It can be seen that the residuals cross the respective thresholds, indicating a faulty sensor. The residuals of the speed and α-axis voltage lie below the threshold. However, the residuals of the β-axis voltage cross the threshold and indicate a fault due to α- and β-axis current faults. Hence, the residuals were further analyzed using various

index-based methods to detect the faulty sensors. The index-based methods the moving root mean square index (MRI), moving-average-based index (MAI), moving-variance-based index (MVI), moving-energy-based index (MEI), first-order-derivative-based index (DBI_1), second-order-derivative-based index (DBI_2), and auto-correlation-based index (A_cI) were designed to detect the faulty sensors in a multi-sensor fault scenario.

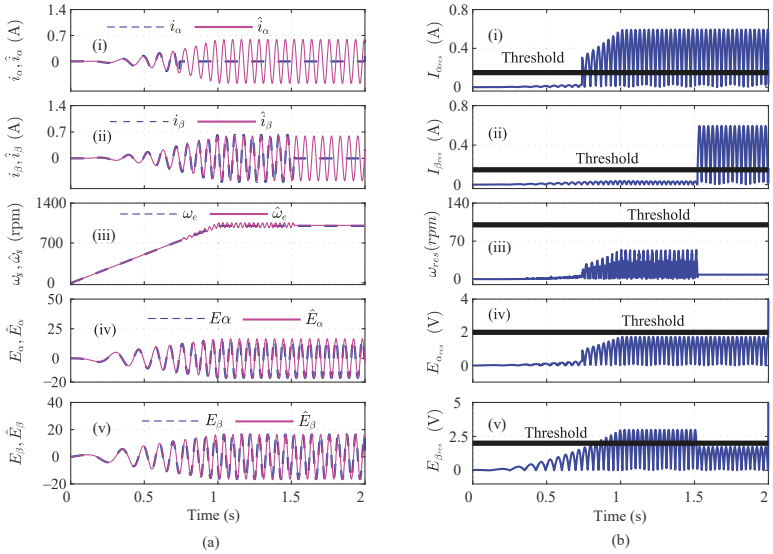


Figure 3. Illustration of current, speed, and voltage sensors' observers during an abrupt i_α and I_β fault scenario at t = 0.739 s and t = 1.52 s; (**a**) actual and estimated signals; (**b**) residuals.

As shown in Figure 4a(i), the MRI of the i_α crosses its threshold and indicates that i_α is faulty at $t = 0.739$. However, the i_α residual lies below the threshold, as shown in Figure 4a(ii). The MAI failed to detect the low-severity current faults in the PMSM. The MVI, MEI, DBI_1, and DBI_2, however, characterized the faulty condition, as shown in Figure 4a(iii), 4a(iv), 4a(v), and 4a(vi), respectively, at $t = 0.740$ s, 0.7402 s, 0.391 s, and 0.7391 s. Using Equation (8), it is clear that $n = 5$ and $n = 1$, clearly indicating that the number of fault detection indices was greater than the indices that failed to detect the faults. From Figure 4, it can be seen that, due to the fault in the α-axis of the stator current, the energies of the $I_{\alpha_{res}}$ and the MRI values were proportionally related, for a constant number of samples in the specified moving window. Hence, it can be seen that after the fault, the residual value increased tremendously, leading to an increase in the energy and, hence, and increase the MRI values also. Similarly, the residuals of i_β were analyzed by using the index-based methods, as shown in Figure 4b. The depicted MRI for the i_β crosses the threshold at $t = 1.525$ s and indicates a faulty i_β sensor, as shown in Figure 4b(i). However, the depicted MAI lies below the threshold, as shown in Figure 4b(ii). The MVI, MEI, DBI_1, and DBI_2 of the i_β residuals lie above the threshold and indicate a faulty i_β sensor, as shown in Figure 4b(iii), 4b(iv), 4b(v), 4b(vi), respectively, at at $t = 1.528$ s, 1.526 s, 1.522 s, and 1.522 s. The AI of i_β also shows that $f = 5$ and $nf = 1$, hence indicating a faulty I_β sensor. Similarly, the indices for the ω_e residual are plotted in Figure 5a. The depicted MRI of the ω touches the threshold slightly, as shown in Figure 5a(i). The indices MAI, MVI, MEI, DBI_1, and DBI_2 are shown in Figure 5a(ii), 5a(iii), 5a(iv), and 5a(v), respectively. The residuals of E_α were also analyzed using the index-based methods of MRI, MAI, MVI, MEI, DBI_1, and DBI_2, respectively as shown in Figure 5b(i), 5b(ii), 5b(iii), 5b(iv), 5b(v), and 5b(vi). The MRI, MAI, MVI, MEI, DBI_1, and DBI_2 for the residuals of the E_β voltage sensors lie below

the threshold, as shown in Figure 6. From the AI analysis, it can be depicted that E_β is fault free.

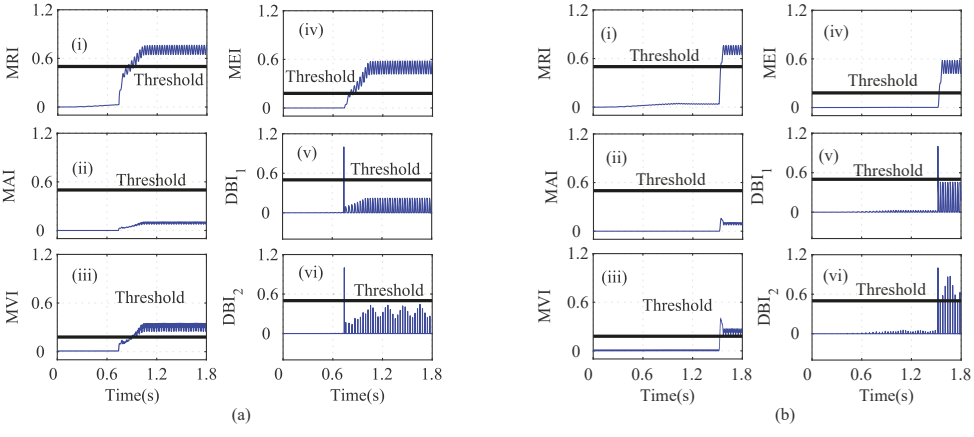


Figure 4. Analysis of residuals for (**a**) I_α and (**b**) I_β using index-based methods ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for the I_α and I_β faults at t = 0.739 s and t = 1.52 .

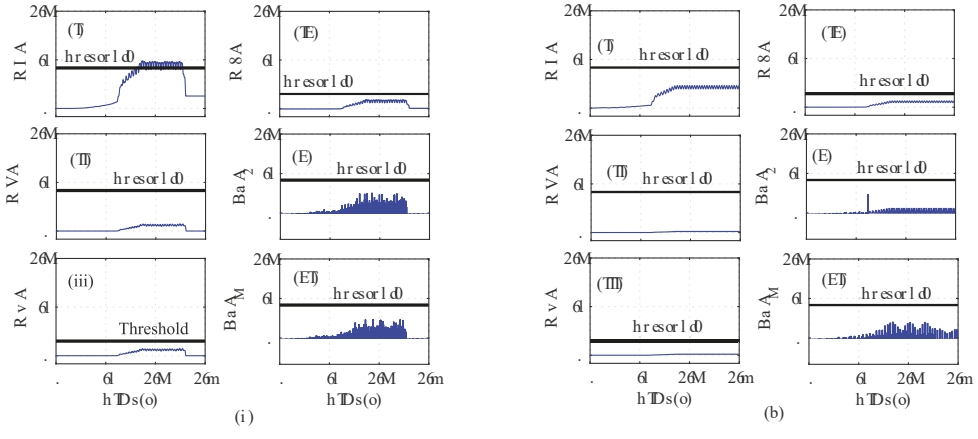


Figure 5. Analysis of residuals for (**a**) ω and (**b**) V_α using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for the I_α and I_β faults at t = 0.739 s and t = 1.52 s.

In the second case, a combination of an incipient and an abrupt fault was also introduced in i_α and E_β at t = 1.20 s and 1.60 s, respectively, as shown in Figure 7a(i) and 7a(v). A load change of 50% was also considered while introducing the i_α and E_β faults. Due to the faults in both sensors, i_β, E_α, and E_β also were affected, as shown in Figure 7a(ii), 7a(iii), and 7a(iv), respectively. The residuals of the i_α current sensor cross the threshold at t = 1.20 s, as shown in Figure 7b(i). Similarly, the residuals of i_β, ω_e, and E_α are shown in Figure 7b(ii), 7b(iii), and 7b(iv), respectively. The E_β residual crosses the threshold and indicates a faulty E_β sensor, as shown in Figure 7b(v). A further analysis was performed to detect the actual faulty sensor by using the index-based analysis methods.

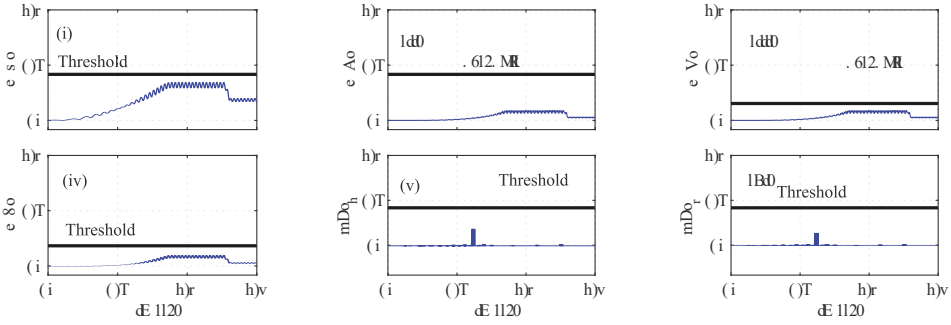


Figure 6. Analysis of residuals for V_β using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for the I_α and I_β fault at t = 0.739 s and t = 1.52 s.

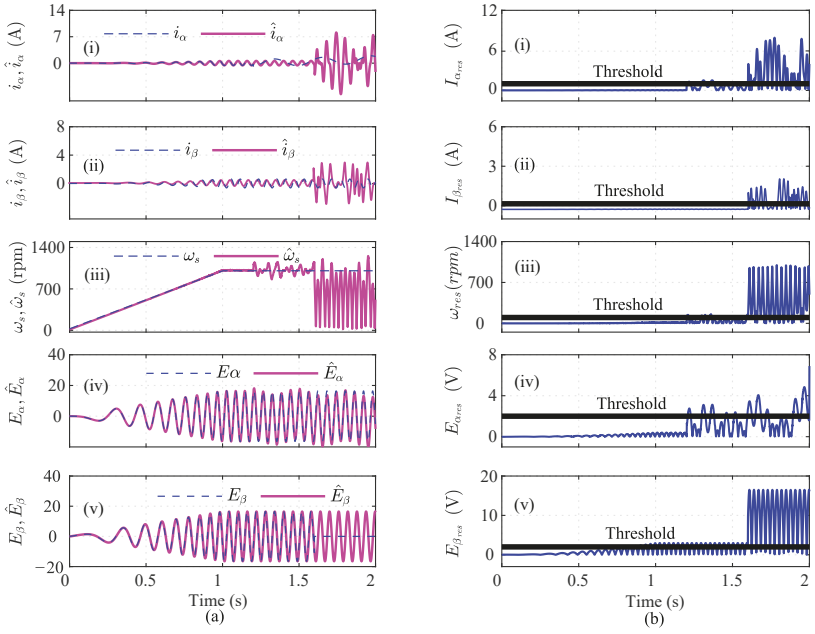


Figure 7. Illustration of current, speed, and voltage sensors' observers during the incipient i_α fault and abrupt V_β fault scenario at t = 1.256 s and t = 1.60 s, respectively; (**a**) actual and estimated signals; (**b**) residuals.

An interturn fault in the i_α sensor can lead to a high fault current in the shorted circuit, which can produce excessive heat and ripples in the torque. The MRI of the i_α sensor indicates a fault at t = 1.258 s, as shown in Figure 8a(i). However, the MAI lies below the threshold, and hence, it was unable to detect the i_α fault, as shown in Figure 8a(ii). The depicted MVI touches the threshold, indicating a fault at t = 1.257 s, as shown in Figure 8a(iii). The MEI, DBI_1, and DBI_2 of the i_α residual increase and cross the threshold at 1.257 s, 1.2563 s, and 1.2563 s, as shown in Figure 8a(iv), 8a(v), and 8a(vi), respectively. The AI was calculated to check the faulty sensor, and it can be seen that $n = 5$ and $nf = 1$; hence, i_α was considered as a faulty sensor. The index-based analysis for the i_β sensor is shown in Figure 8b. The indices (MRI, MAI, MVI, MEI, DBI_1, and DBI_2 lie below the selected threshold, as shown in Figure 8b(i), 8b(ii), 8b(iii), 8b(iv), 8b(v), and 8b(vi)), respectively.

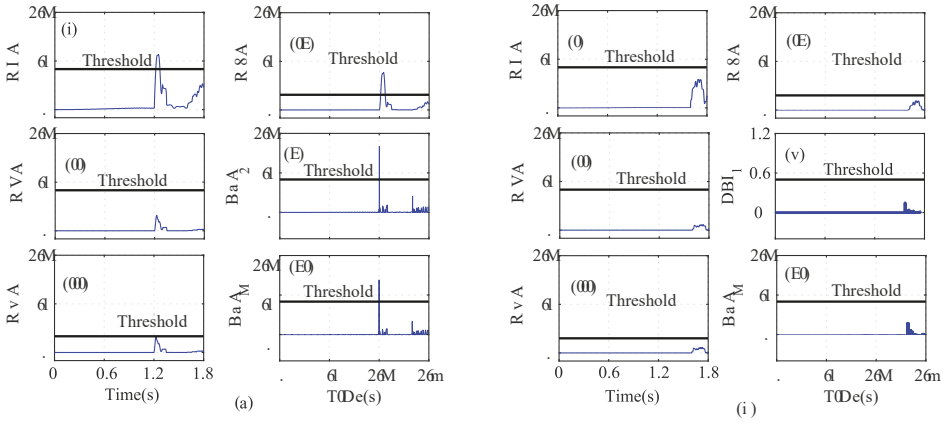


Figure 8. Analysis of residuals for (**a**) I_α and (**b**) I_β using index-based methods ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) during the incipient i_α fault and abrupt V_β fault scenario at t = 1.256 s and t = 1.60 s, respectively.

The speed (ω_s) residual was also analyzed by using the index-based methods, as shown in Figure 9a. The MRI, MAI, MVI, MEI, DBI_1, and DBI_2 lie below the threshold, as shown in Figure 9a(i), 9a(ii), 9a(iii), 9a(iv), 9a(v), and 9a(vi), respectively. The analysis of the E_α residual was also performed using the index-based methods. The depicted MRI, MAI, MVI, MEI, DBI_1, and DBI_2 lie below the selected threshold, which indicates that E_β is non-faulty, as shown in Figure 9b(i), 9b(ii), 9b(iii), 9b(iv), 9b(v), 9b(vi), respectively.

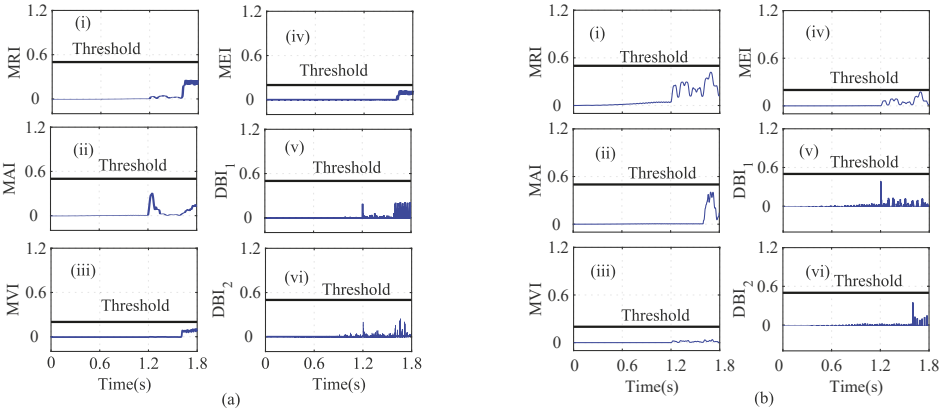


Figure 9. Analysis of residuals for (**a**) W and (**b**) V_α using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for incipient I_α fault and abrupt V_β fault at t = 1.256 s and t = 1.60 s.

However, the calculated MRI for the E_β residual surpasses the threshold at t = 1.60 s, indicating a faulty sensor, as shown in Figure 10i. The MAI and MVI, however, stay below the threshold and discriminate the change as a healthy E_β sensor, as shown in Figure 10ii and 10iii, respectively. However, as shown in Figure 10iv, the MEI for the E_β residual slightly exceeds the threshold at t = 1.62 s and indicates that E_β is faulty. The DBI_1 and DBI_2 of the E_β sensor residual also increase and cross the threshold at 1.605 s and 1.605 s, as shown in Figure 10v and 10vi, respectively. The AI was thus calculated to further analyze the index-based methods. It can be seen that $n = 4$ and $nf = 2$; hence, E_β was considered as a faulty sensor.

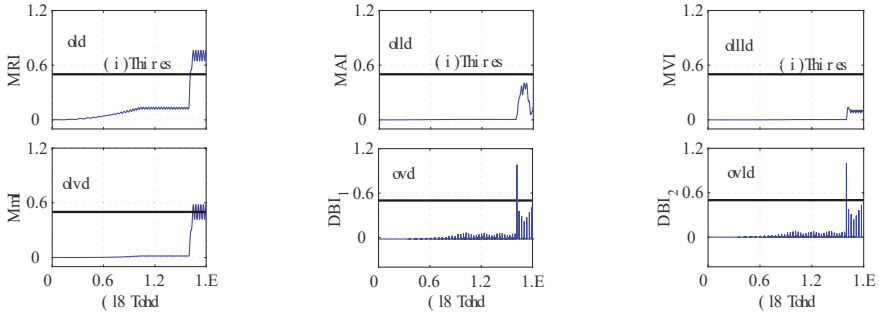


Figure 10. Analysis of residuals for V_β using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for incipient I_α fault and abrupt V_β fault at t = 1.256 s and t = 1.60 s.

In the third case, the effect of the speed sensor fault in addition to the current sensor fault was analyzed for accurate fault detections. In this regard, an incipient fault was introduced in i_α sensor and an abrupt fault in ω_e sensor at t = 1.25 s and t = 1.50 s, respectively, as shown in Figure 11. A random noise of 20% was also introduced in the i_α sensor. The actual and the estimated states are shown in Figure 11a, and the residuals are shown in Figure 11b, respectively. The MRI, MVI, MEI, DBI_1, and DBI_2 lie above the threshold and indicate a faulty sensor at 1.26 s, 1.28 s, 1.48 s, 1.253 s, and 1.253 s, as shown in Figure 12a(i), Figure 12a(iii), Figure 12a(iv), Figure 12a(v), and Figure 12a(vi), respectively. However, the MAI values lie below the threshold, as shown in Figure 12a(ii). Furthermore, it can be seen that $f = 5$ and $nf = 1$. Hence, the i_α sensor was concluded to be faulty. Similarly for the i_β index analysis, all the indices, the MRI, MAI, MVI, MEI, DBI_1, and DBI_2, lie below the threshold, as shown in Figure 12b. Hence, i_β is not faulty. The analysis of ω was also performed in a similar manner.

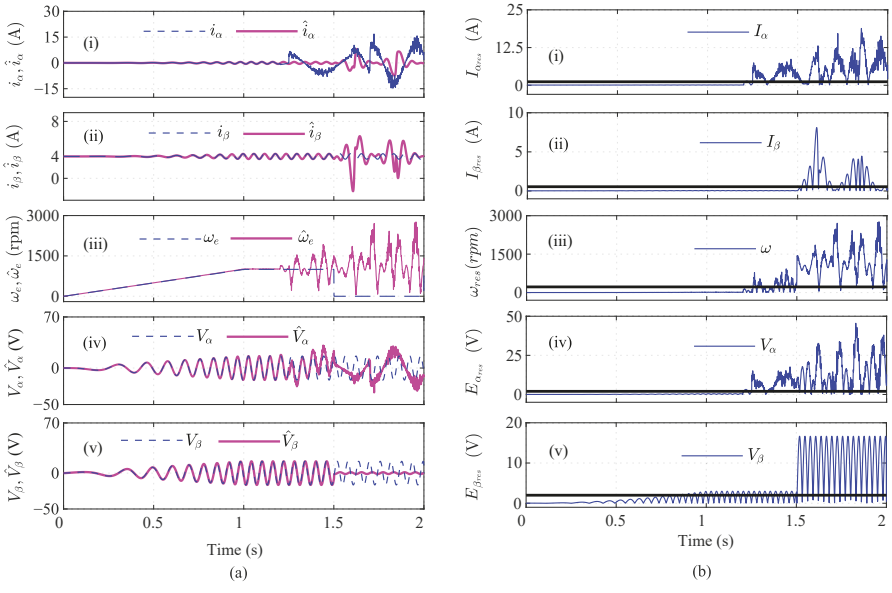


Figure 11. Illustration of current, speed, and voltage sensors' observers during the 20% noisy incipient i_α fault and abrupt W_e fault scenario at t = 1.25 s and t = 1.50 s, respectively; (**a**) actual and estimated signals; (**b**) residuals.

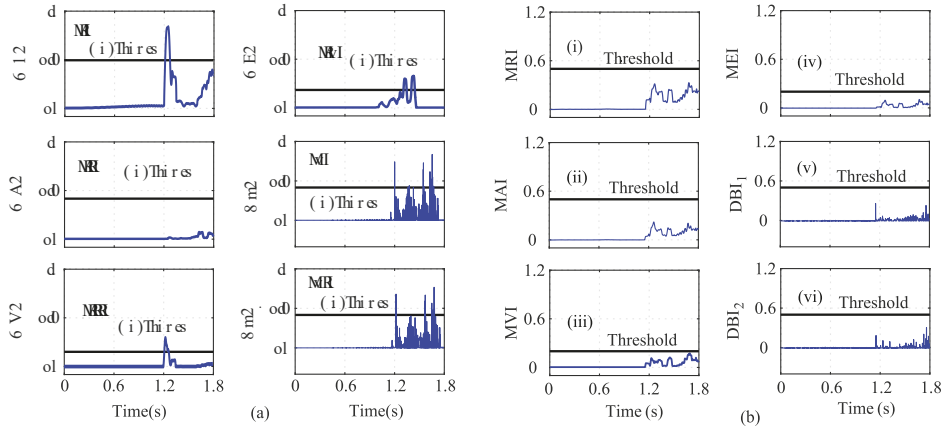


Figure 12. Analysis of residuals for (**a**) i_α and (**b**) i_β using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for the 20% noisy incipient i_α fault and abrupt W_e fault scenario at t = 1.25 s and t = 1.50 s, respectively.

As shown in Figure 13a, it can be seen that the MRI, MAI, MVI, MEI, DBI_1, and DBI_2 cross the threshold, and hence, the abrupt ω fault can be detected, as shown in Figure 13a(i–vi). However, in the case of index analysis method application for the E_α sensor, the MRI showed a slow increase in its values and slightly touches the threshold, as shown in Figure 13b(i). The other indices still remain below the threshold. Further utilizing the AI analysis, it can be seen $f = 1$ and $nf = 5$, which implies that $f < nf$; hence E_β is non-faulty. Similarly, the index method for E_β sensor analysis is shown in Figure 14. The calculated AI shows that $f = 1$ and $nf = 5$, and hence, $f < nf$ indicates that E_β is fault free.

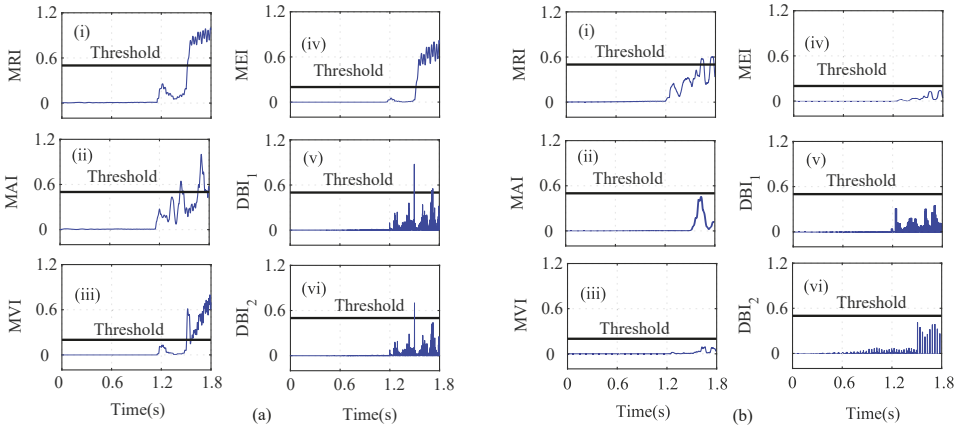


Figure 13. Analysis of residuals for (**a**) ω_e and (**b**) E_α using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for the 20% noisy incipient i_α fault and abrupt ω_e fault scenario at t = 1.25 s and t = 1.50 s, respectively.

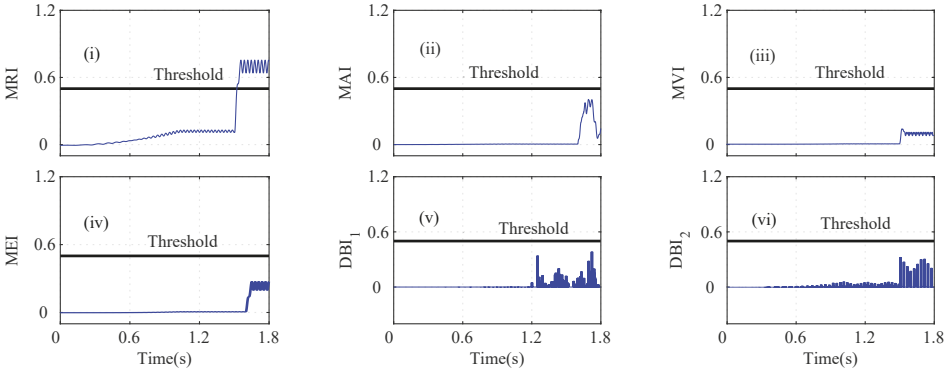


Figure 14. Analysis of residuals for E_β using index-based method ((i) MRI, (ii) MAI, (iii) MVI, (iv) MEI, (v) DBI_1 and (vi) DBI_2) for the 20% noisy incipient i_α fault and abrupt ω_e fault scenario at t = 1.25 s and t = 1.50 s, respectively.

To show the efficacy of the proposed method, an incipient fault in i_α sensor was considered at $t = 1.25$ s, and and intermittent fault was considered in the ω_e sensor, with the first fault occurring at $t = 0.5$ s and ending at $t = 0.8$ s. The second intermittent fault occurred at $t = 1.5$ s and ended at $t = 1.8$ s, as shown in Figure 15a. The illustration of the HOSM observer for the i_α and ω_e faults is shown in Figure 15b. The A_cI-based index was calculated to analyze the residuals of the faulty sensors. Using this A_cI-based method, it can be seen that the incipient fault in the i_α sensor failed to be detected, as shown in Figure 16i, whereas the intermittent faults in the ω_e sensor were detected with a delay, as shown in Figure 16iii.

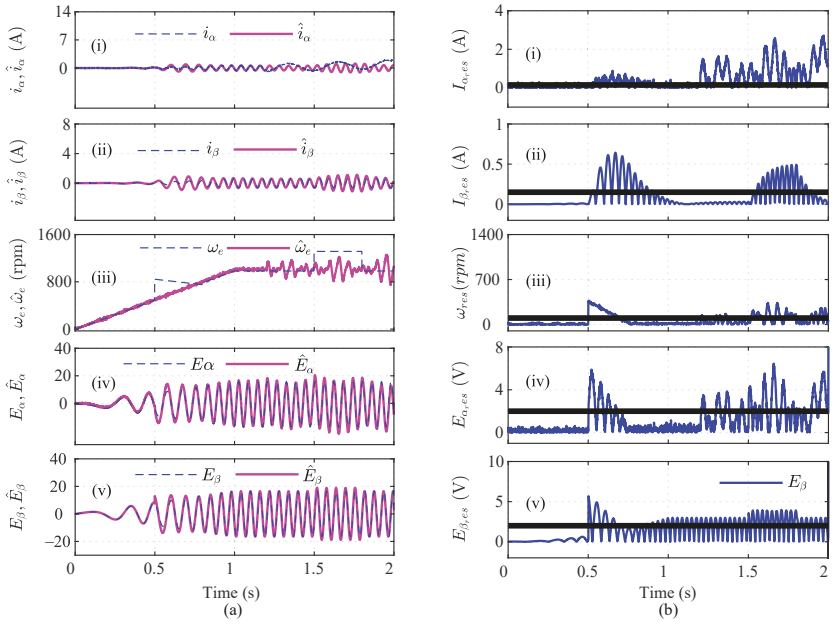


Figure 15. Illustration of current, speed, and voltage sensors' observers during the incipient i_α fault at t = 1.25 s and intermittent W_ω fault scenario at t = 0.5 s and t = 1.50 s, respectively; (**a**) actual and estimated signal; (**b**) generated residuals.

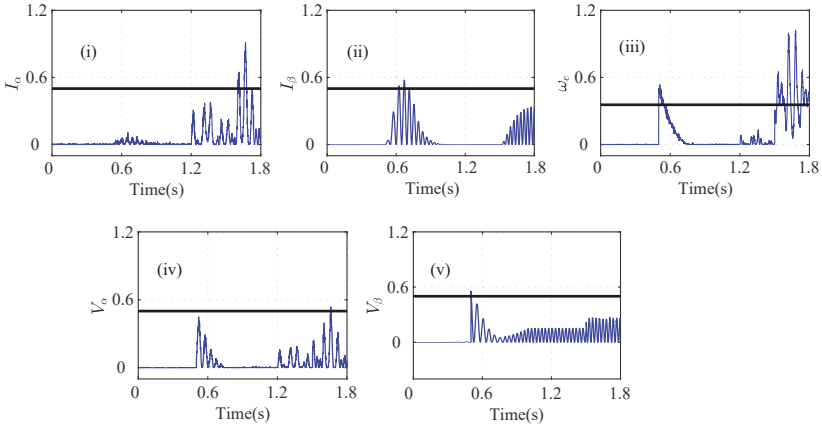


Figure 16. A_cI-based analysis for current ((i) I_α, (ii) I_β), speed ((iii) ω_e), and voltage sensors' ((iv) V_α, (v) V_β) residuals during the incipient i_α fault at t = 1.25 s and intermittent W_ω fault scenario at t = 0.5 s and t = 1.50 s, respectively.

5. Discussion

In this research, we took into account both sudden and incipient sensor faults in industrial drives in order to perform multi-sensor fault detection. Different conditions were considered by sequentially introducing the faults in each sensor. Indices (MRI, MAI, MVI, MEI, DBI_1, and DBI_2) were designed to detect the faults by selecting a particular threshold. The threshold for the indices was designed by using Otsu-based thresholding techniques.

Different cases such as low-severity current faults, sudden speed changes, and changes in the load were considered on the basis of index analysis for detecting the single- and multi-sensor faults. As shown in Figure 4a, during low-severity i_α and i_β faults, the MRI, MVI, and MEI detected the fault after a certain delay. However, the indices DBI_1 and DBI_2 detected the faulty i_α sensor with a minimum delay compared to the other indices. Similarly, in Figure 4b, the MRI, MVI, and MEI experienced a certain delay in detecting the i_β fault. On the other hand, DBI_1 and DBI_2 detected the faulty i_β sensor immediately after the fault's occurrence. The indices for ω_e, E_α, and E_β did not show any sudden change and indicated fault-free sensor data, by holding the property of the auxiliary index.

In Figure 8a, due to the load change in the drive, the MRI, MVI, MEI, DBI_1, and DBI_2 detected the fault of the i_α sensor data. The indices DBI_1 and DBI_2, however, detected the fault with a minimum delay compared to the other indices. The E_β sensor fault in Figure 8 also shows a variation and crosses the respective thresholds when the MRI, MEI, DBI_1, and DBI_2 were utilized. The auxiliary index (AI) plays a vital role in selecting the accurate fault without discriminating the outputs of the indices.

A random noise of 20% was also introduced in the i_α sensor, as shown in Figure 11a(i). Due to noise in the sensors, the MRI value detected the i_α sensor fault at $t = 1.26$ s. The MVI also detected the fault at $t = 1.30$ s with the maximum delay. The DBI_1 and DBI_2 detected the noisy i_α sensor at $t = 1.26$ s and $t = 1.262$ s, respectively. As shown in Figure 13, all seven indices, MRI, MAI, MVI, MEI, DBI_1, DBI_2, and A_cI, detected the abrupt speed (ω_e) fault. As shown in Figure 17, the accuracies of the indices were calculated, and it can be seen that the A_cI had the lowest accuracy for various fault conditions.

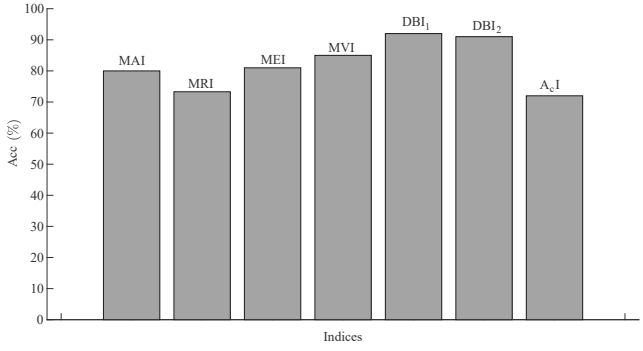


Figure 17. Accuracies of the indices used in the proposed method.

Similarly, the dependability of the indices was calculated for the comparison of all the indices under various conditions, as shown in Figure 18.

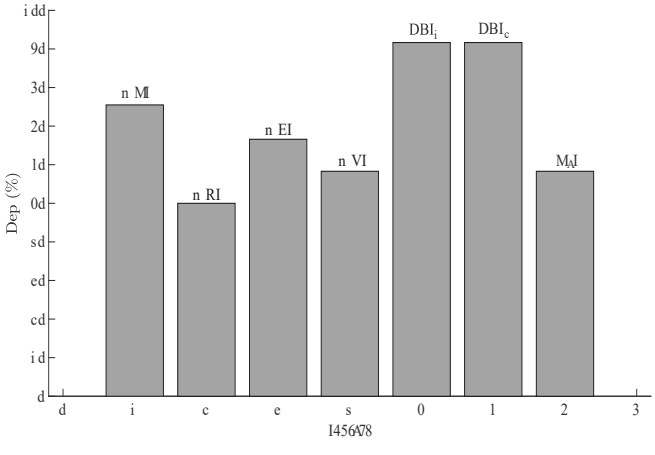


Figure 18. Dependability of the indices used in the proposed method.

The indices were calculated for all five sensors under three different cases mainly, low-severity current faults, sensor faults during sudden load changes, and impacts of sensor speed changes along with other sensors. A random noise of more than 20% was also tested in the sensors for multiple fault detections. However, due to increase in the noise, the MRI increased and crossed the threshold, indicating a false faulty sensor. The proposed index method can be improved by using a low-pass filter for signals with noises greater than 20%, as the DBI_1 and DBI_2 became more sensitive to noise, compromising the fast detection property compared with other indices. Hence, in this case, adaptive thresholds can also be incorporated to prevent the system from false faulty sensor signal detection.

6. Conclusions

In this work, different types of sensor faults were analyzed for fault detection by using multiple index-based methods for a healthy drive. Seven index-based methods were analyzed for the identification of the changes that occurred in the faulty sensors and the non-faulty sensors. The results showed that the MRI, MEI, MVI, DBI_1, and DBI_2 were able to detect the low-severity faults. The combination of both incipient current and an abrupt voltage fault during the load change could be detected by the MRI, MEI, DBI_1, and DBI_2 accurately. The seven index-based methods could also be used to identify variations in the speed sensor when they were combined with a fault in the current sensor. A combination

of an intermittent fault in the speed sensor and an incipient fault in the beta-axis of the current sensor was also simulated, and the index-based methods were able to identify the faulty sensors. The simulated results conducted on various fault scenarios showed that index-based analysis can be employed for fast fault detection. Future work will focus on experimental validation of the proposed method on a PMSM motor.

Author Contributions: Conceptualization, K.C.V.; methodology, D.N.; software, D.N.; validation, K.C.V. and D.N.; formal analysis, D.N.; investigation, K.C.V.; resources, K.C.V.; writing—original draft preparation, D.N.; writing—review and editing, K.C.V.; supervision, K.C.V.; project administration, K.C.V.; funding acquisition, K.C.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A4A1023248).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tao, H.; Peng, T.; Yang, C.; Gao, J.; Chen, Z.; Yang, C.; Gui, W. An FCS-MPC-based open-circuit and current sensor fault diagnosis method for traction inverters with two current sensors. *Int. J. Electr. Power Energy Syst.* **2023**, *144*, 108526. [CrossRef]
2. Chen, Y.; Gou, L.; Li, H.; Wang, J. Sensor Fault Diagnosis of Aero Engine Control System Based on Honey Badger Optimizer. *IFAC-PapersOnLine* **2022**, *55*, 228–233. [CrossRef]
3. Long, Z.; Zhang, X.; Zhang, L.; Qin, G.; Huang, S.; Song, D.; Shao, H.; Wu, G. Motor fault diagnosis using attention mechanism and improved adaboost driven by multi-sensor information. *Measurement* **2021**, *170*, 108718. [CrossRef]
4. Frisk, E.; Jarmolowitz, F.; Jung, D.; Krysander, M. Fault Diagnosis Using Data, Models, or Both–An Electrical Motor Use-Case. *IFAC-PapersOnLine* **2022**, *55*, 533–538. [CrossRef]
5. Zhao, Z.; Liu, P.X.; Gao, J. Model-based fault diagnosis methods for systems with stochastic process—A survey. *Neurocomputing* **2022**, *513*, 137–152. [CrossRef]
6. Man, J.; Dong, H.; Jia, L.; Qin, Y. AttGGCN Model: A Novel Multi-Sensor Fault Diagnosis Method for High-Speed Train Bogie. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19511–19522. [CrossRef]
7. Goktas, T. Evaluation and Classification of Double Bar Breakages through Three-Axes Vibration Sensor in Induction Motors. *IEEE Sens. J.* **2022**, *22*, 13602–13611. [CrossRef]
8. Ma, S.; Yuan, Y.; Wu, J.; Jiang, Y.; Jia, B.; Li, W. Multisensor decision approach for HVCB fault detection based on the vibration information. *IEEE Sens. J.* **2020**, *21*, 985–994. [CrossRef]
9. Cao, X.; Xu, X.; Duan, Y.; Yang, X. Health Status Recognition of Rotating Machinery Based on Deep Residual Shrinkage Network under Time-varying Conditions. *IEEE Sens. J.* **2022**, *22*, 18332–18348. [CrossRef]
10. Furse, C.M.; Kafal, M.; Razzaghi, R.; Shin, Y.J. Fault diagnosis for electrical systems and power networks: A review. *IEEE Sens. J.* **2020**, *21*, 888–906. [CrossRef]
11. Piltan, F.; Kim, J.M. Bearing fault diagnosis by a robust higher-order super-twisting sliding mode observer. *Sensors* **2018**, *18*, 1128. [CrossRef]
12. Satpathi, K.; Yeap, Y.M.; Ukil, A.; Geddada, N. Short-time Fourier transform based transient analysis of VSC interfaced point-to-point DC system. *IEEE Trans. Ind. Electron.* **2017**, *65*, 4080–4091. [CrossRef]
13. Pang, B.; Tang, G.; Tian, T. Rolling bearing fault diagnosis based on SVDP-based kurtogram and Iterative autocorrelation of Teager energy operator. *IEEE Access* **2019**, *7*, 77222–77237. [CrossRef]
14. He, M.; He, D. Deep learning based approach for bearing fault diagnosis. *IEEE Trans. Ind. Appl.* **2017**, *53*, 3057–3065. [CrossRef]
15. Kommuri, S.K.; Lee, S.B.; Veluvolu, K.C. Robust sensors-fault-tolerance with sliding mode estimation and control for PMSM drives. *IEEE/ASME Trans. Mechatron.* **2017**, *23*, 17–28. [CrossRef]
16. Andersson, A.; Thiringer, T. Motion sensorless IPMSM control using linear moving horizon estimation with Luenberger observer state feedback. *IEEE Trans. Transp.* **2018**, *4*, 464–473. [CrossRef]
17. Mazzoletti, M.A.; Bossio, G.R.; De Angelo, C.H.; Espinoza-Trejo, D.R. A model-based strategy for interturn short-circuit fault diagnosis in PMSM. *IEEE Trans. Ind. Electron.* **2017**, *64*, 7218–7228. [CrossRef]
18. Gou, L.; Shen, Y.; Zheng, H.; Zeng, X. Multi-fault diagnosis of an aero-engine control system using joint sliding mode observers. *IEEE Access* **2020**, *8*, 10186–10197. [CrossRef]
19. Rigatos, G.; Abbaszadeh, M. Fault Diagnosis for a PDE Suspended-Bridge Model With Kalman Filter and Statistical Decision Making. *IEEE Syst. J.* **2020**, *15*, 2137–2147. [CrossRef]

20. Zou, S.; Zhao, W.; Liang, W.; Wang, C.; Chen, F. Fault Diagnosis and Fault-Tolerant Compensation Strategy for Wheel Angle Sensor of Steer-by-Wire Vehicle via Extended Kalman Filter. *IEEE Sens. J.* **2021**, *22*, 1756–1766. [CrossRef]
21. Choi, K.; Kim, Y.; Kim, S.K.; Kim, K.S. Current and position sensor fault diagnosis algorithm for PMSM drives based on robust state observer. *IEEE Trans. Ind. Electron.* **2020**, *68*, 5227–5236. [CrossRef]
22. Salehifar, M.; Salehi Arashloo, R.; Moreno-Eguilaz, M.; Sala, V.; Romeral, L. Observer-based open transistor fault diagnosis and fault-tolerant control of five-phase permanent magnet motor drive for application in electric vehicles. *IET Power Electron.* **2015**, *8*, 76–87. [CrossRef]
23. Wu, Y.; Zhao, D.; Liu, S.; Li, Y. Fault detection for linear discrete time-varying systems with multiplicative noise based on parity space method. *ISA Trans.* **2022**, *121*, 156–170. [CrossRef] [PubMed]
24. Allal, A.; Khechekhouche, A. Diagnosis of induction motor faults using the motor current normalized residual harmonic analysis method. *Int. J. Electr. Power Energy Syst.* **2022**, *141*, 108219. [CrossRef]
25. Konar, P.; Chattopadhyay, P. Multi-class fault diagnosis of induction motor using Hilbert and Wavelet Transform. *Appl. Soft Comput.* **2015**, *30*, 341–352. [CrossRef]
26. Capolino, G.A.; Antonino-Daviu, J.A.; Riera-Guasp, M. Modern diagnostics techniques for electrical machines, power electronics, and drives. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1738–1745. [CrossRef]
27. Wu, H.; Feng, J. A Review of Fault Diagnosis Methods of Robot Joint Servo System. In Proceedings of the 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 13–16 October 2020; pp. 61–66.
28. Ismail, A.; Saidi, L.; Sayadi, M. Wind turbine power converter fault diagnosis using DC-link voltage time–frequency analysis. *Wind Eng.* **2019**, *43*, 329–343. [CrossRef]
29. Li, Y.; Wang, Y.; Zhang, Y.; Zhang, J. Diagnosis of inter-turn short circuit of permanent magnet synchronous motor based on deep learning and small fault samples. *Neurocomputing* **2021**, *442*, 348–358. [CrossRef]
30. Jlassi, I.; Cardoso, A.J.M. A single method for multiple IGBT, current, and speed sensor faults diagnosis in regenerative PMSM drives. *IEEE Trans. Emerg. Sel. Top. Power Electron.* **2019**, *8*, 2583–2599. [CrossRef]
31. Li, M.; Lee, K.M.; Hanson, E. Sensor Fusion Based on Embedded Measurements for Real-Time Three-DOF Orientation Motion Estimation of a Weight-Compensated Spherical Motor. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 9508009. [CrossRef]
32. Zhu, M.; Hu, W.; Kar, N.C. Multi-sensor fusion-based permanent magnet demagnetization detection in permanent magnet synchronous machines. *IEEE Trans. Mag.* **2018**, *54*, 8110106. [CrossRef]
33. Tran, C.D.; Palacky, P.; Kuchar, M.; Brandstetter, P.; Dinh, B.H. Current and speed sensor fault diagnosis method applied to induction motor drive. *IEEE Access* **2021**, *9*, 38660–38672. [CrossRef]
34. Yu, M.; Wang, D. Model-based health monitoring for a vehicle steering system with multiple faults of unknown types. *IEEE Trans. Ind. Electron.* **2013**, *61*, 3574–3586.
35. Niu, Y.; Gao, M.; Sheng, L. Fault-tolerant state estimation for stochastic systems over sensor networks with intermittent sensor faults. *Appl. Math. Comput.* **2022**, *416*, 126723. [CrossRef]
36. Goh, T.Y.; Basah, S.N.; Yazid, H.; Safar, M.J.A.; Saad, F.S.A. Performance analysis of image thresholding: Otsu technique. *Measurement* **2018**, *114*, 298–307. [CrossRef]
37. Kommuri, S.K.; Defoort, M.; Karimi, H.R.; Veluvolu, K.C. A robust observer-based sensor fault-tolerant control for PMSM in electric vehicles. *IEEE Trans. Ind. Electron.* **2016**, *63*, 7671–7681. [CrossRef]
38. Xu, D.; Ding, B.; Jiang, B.; Yang, W.; Shi, P. Nonsingular fast terminal sliding mode control for permanent magnet linear synchronous motor via high-order super-twisting observer. *IEEE/ASME Trans. Mechatron.* **2021**, *27*, 1651–1659. [CrossRef]
39. Bernard, P.; Andrieu, V. Luenberger observers for nonautonomous nonlinear systems. *IEEE Trans. Autom. Control* **2018**, *64*, 270–281. [CrossRef]

sensors

MDPI

Article

Cluster Migration Distance for Performance Degradation Assessment of Water Pump Bearings

Zhongping Zhai [1], Zihao Zhu [2], Yifan Xu [2], Xinhang Zhao [2], Fang Liu [2,*] and Zhihua Feng [1,*]

[1] Department of Precision Mechanics and Precision Instruments, University of Science and Technology of China, Hefei 230027, China
[2] School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China
* Correspondence: liufang1@mail.ustc.edu.cn (F.L.); fff@ustc.edu.cn (Z.F.)

Abstract: Because the signal of water pump bearing is seriously disturbed by noise and the fault evolution is complex, it is difficult to describe the performance degradation trend of water pump bearing in a timely and accurate manner using the traditional performance degradation index (PDI). In this paper, a new Cluster Migration Distance (CMD) algorithm is proposed. The extraction of the indicator includes the following four steps: First, the relevant blind separation is used to extract the useful signal of the monitored bearing from the mixed signal; secondly, the impact component is further enhanced by wavelet packet analysis. Then, the redundancy of the original feature vectors is eliminated using our previously proposed KJADE (Kernel Joint Approximate Diagonalization of Eigen-matrices) method. Finally, the newly proposed CMD index is computed as PDI. By calculating the offset trajectory of the feature cluster centroid in the continuous running process of the bearing, CMD can aptly deal with the complex and variable features in the fault evolution process of the water pump bearing. The whole-life monitoring data of a 220 KW water pump system are processed. The results show that the proposed CMD index has better early-warning ability and monotonicity than the traditional kurtosis index.

Keywords: water pump bearing; fault diagnosis; blind source separation; feature fusion; feature distance; dynamic center of mass

Citation: Zhai, Z.; Zhu, Z.; Xu, Y.; Zhao, X.; Liu, F.; Feng, Z. Cluster Migration Distance for Performance Degradation Assessment of Water Pump Bearings. *Sensors* **2022**, *22*, 6809. https://doi.org/10.3390/s22186809

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 16 August 2022
Accepted: 5 September 2022
Published: 8 September 2022

1. Introduction

Water pumps play an important role in urban drinking water treatment, agricultural irrigation, and other scenarios. As one of the key components of water pumps, bearings are of great significance to ensure their healthy and reliable operation. However, due to their high load and high-speed operation over a long duration, various kinds of faults inevitably occur [1–3].

Dutta N. et al. [4]. proposed an artificial neural network-based model for the diagnosis of water pump bearings with higher accuracy than other machine-learning algorithms. Tang Jing et al. [5] converted the collected vibration signals into frequency domain signals and used frequency domain features and support vector machines to diagnose the faults of water pump bearings. Li Tao et al. [6,7] proposed an adaptive algorithm based on a particle swarm optimization algorithm using the convolutional neural network fault diagnosis method. Other research [8] proposed the method of the wavelet packet to denoise the vibration signal of the centrifugal pump bearing and extracted the frequency band energy representing the corresponding bearing fault and used a BP neural network for training and fault identification. Han Hui et al. [9] proposed a water pump bearing diagnosis method based on stack noise reduction and self-encoding, which obtained a more robust feature representation and could precisely diagnose the fault state of water pump bearings. Bie Zhongzheng et al. [10] introduced the principle of peak energy analysis and used the peak energy value to judge the failure degree of the water pump bearing.

Denoising and accurate performance degradation index extraction are two major challenges in achieving effective performance degradation assessment of water pump bearings. In practical engineering applications, the water pump is usually driven to rotate by a motor through a coupling, and four bearings are usually installed coaxially on the main shaft. Therefore, the vibration signal of each bearing will receive vibration interference from other bearings, which introduces difficulties in the accurate extraction of the weak fault information in the early stages of the bearing. Therefore, it is necessary to study effective means to effectively eliminate noise. Bearing failures, on the other hand, have complex evolutionary processes, often starting with one failure type, such as localized material tribes, and then more complex composite failures may emerge as bearing components interact with localized defects. Therefore, accurately characterizing the performance degradation trend of water pump bearings is one of the main challenges.

Regarding research on noise reduction, Pinlu et al., proposed a deep learning method for efficient noise reduction and feature extraction, which is based on a combination of residual construction units, soft thresholding, and global context. However, this method has the disadvantage of a shared threshold [11]. Sandaram Buchaiah and Piyush Shakya et al., applied different signal processing techniques to extract 72 raw features from vibration data collected experimentally on bearings. They used a random forest method to select a subset of relevant features from the extracted features. The selected features were fused through a 14-dimensional dimensionality-reduction technique, from which two-dimensional relevant performance indicators are extracted and compared between techniques to determine the most effective technique. However, these algorithms are numerous and they need to be compared when they are used, and it is impossible to determine which one is the optimal algorithm [12]. Fei-Ping Du et al., proposed an adaptive regularization parameter selection strategy to denoise vibration signals using a sparse redundant representation model. However, this posterior method requires a great deal of computation of the FP metric, and although this method is very effective, it has not been applied to all scenarios [13].

Regarding research on performance degradation indicators, Yang Chuang Yan et al., established a new RUL prediction model for the problems of poor residual life prediction performance and the single performance degradation index of rolling bearings, and the method showed a good performance in prediction accuracy and reliability [14]. Yan Xiao li et al., proposed a bearing performance degradation evaluation algorithm based on CMMP and feature fusion. Mathematical morphological operations are driven by partial differential equations (PDEs) for the accurate assessment of bearing life cycle failure datasets [15]. Tao Zan et al., introduced joint approximate diagonalization of a characteristic matrix (JADE) and a particle swarm optimization support vector machine (PSO–SVM) in the prediction of the performance degradation trend of rolling bearings, and realized the performance degradation trend and residual performance of rolling bearings under small samples, with an accurate prediction of service life [16].

In the above studies, it is generally considered that only one type of failure occurs in the bearing during the degradation process, but in fact, in actual engineering cases, we often find that there are composite failures. This is because, after the early failure of a bearing, the interaction of the local failure with the bearing components can lead to more complex compound failures. At this time, some traditional performance degradation indicators, such as kurtosis, will fluctuate significantly, which affects the accurate judgment of the performance degradation trend. Therefore, in this study, we innovatively propose a novel performance degradation metric extraction method called Cluster Migration Distance (CMD). CMD does not focus on the growth of features, but rather on the change in features, becoming lower or higher. Degradation trends are assessed by tracking changes in features. At the same time, in order to solve the problem of strong noise interference, we comprehensively use the non-target vibration source interference elimination ability of blind separation and the impact component identification ability of wavelet analysis to achieve noise elimination and use our specially proposed KJADE method for feature fusion processing. In order to verify the effectiveness of the proposed method, we verify

the life-cycle monitoring data of a 220 KW water pump system in practical engineering applications. The results show that the early warning ability and the monotonicity of the proposed CMD index are better than the traditional kurtosis indicators.

The details of this work are described in the following sections. Section 2 provides the detailed flow of the proposed approach. Section 3 presents the analysis results and related discussions of an engineering application case. Section 4 concludes this work and points out future research work.

2. Method and Process

In practical engineering applications, the early fault signal of a water pump bearing is seriously interfered with, and the fault evolution process is complex and changeable. For the purpose of solving this problem, this paper proposes a new performance degradation index extraction method called the Cluster Migration Distance (CMD). At the same time, blind separation, wavelet packet analysis, and the KJADE feature fusion method are comprehensively used to assist in the accurate extraction of bearing performance degradation indicators.

The proposed method mainly includes the following four steps: (1) In order to eliminate the interference of the coaxial vibration source in the pump system, decorrelated blind separation is used to separate the signal of the monitored bearing from the acquired signals of multiple measurement points. (2) Wavelet packet analysis is used to further enhance the shock components closely related to the diagnostic information from the signal; the comprehensive application of the blind separation and wavelet packet analysis methods can effectively eliminate noise interference. (3) Our proposed KJADE method is used to eliminate redundancy and feature fusion on the original time–frequency domain feature vectors; the KJADE method combines the nonlinear processing capability of the kernel method and the advantages of high-order cumulant calculation in the JADE method, which can effectively extract features that characterize fault information. (4) Lastly, the CMD index is calculated to be used as a final indicator to describe the degradation trend of the monitored bearing performance.

The schematic flow chart of the method and the detailed flow chart of the algorithm are shown in Figures 1 and 2, respectively, and the details of the method will be described below.

2.1. Vibration Source Interference Cancellation Based on Blind Separation

Principal Component Analysis (PCA) [13,16] is a multidimensional data analysis method commonly used in statistical analysis, and it is able to find implicit statistical features from raw data. It is very effective in data dimensionality reduction [17], information compression, and de-correlation between data. In this section, PCA is used to reduce the dimensionality of the two-dimensional signals collected by the horizontal and vertical sensors of the same measuring point before blind separation and extract the main feature information. We assume that the vibration source signal of the monitored bearing is $s(t)$, and other coaxial vibration source signals are $s_i(t)$. After the pump mixing system, the random vector formed by the observation signals of the same measuring point is $x = (x_{i-1}, x_{i-2})^T$, (i = 1, . . . , N), N is the number of measuring points, and its mean $m_x = E(x) = 0$. We then find an orthogonal transformation matrix $\boldsymbol{W}$ and perform an orthogonal transformation on the random vector, so that the random variables in the output random vector $y = \boldsymbol{W}x$ are not correlated with each other.

$$\boldsymbol{C}_y = E\left\{yy^T\right\} = diag(\lambda_1, \lambda_2) \tag{1}$$

The orthogonal matrix W is obtained by decomposing the eigenvalues of the covariance matrix $\boldsymbol{C}_x$ of x. Usually, $\boldsymbol{C}_x$ is a real symmetric matrix, decomposed as:

$$\boldsymbol{C}_x = U\Lambda U^T \tag{2}$$

where $U = (u_1, u_2)$, $\Lambda = diag(\lambda_1, \lambda_2)$ and u_1 and u_2 are the eigenvectors of the covariance matrix C_x. The eigenvectors are orthogonal to each other, that is $E\{{u_i}^T u_j\} = 0 (i \neq j \in 1, 2)$, and λ_i is the corresponding eigenvalue and is presented as:

$$\lambda_i = E\left\{\left({u_i}^T x\right)^2\right\} > \lambda_j = E\left\{\left({u_j}^T x\right)^2\right\}, i < j \in 1, 2 \tag{3}$$

It can be seen that PCA establishes a set of orthogonal bases u_1, u_2 for the 2-dimensional data space, and u_1 is orthogonal to $u_2 . y_i = {u_i}^T x, i = 1, 2$. We then take the first $y_i(t)$ after dimensionality reduction as the principal component.

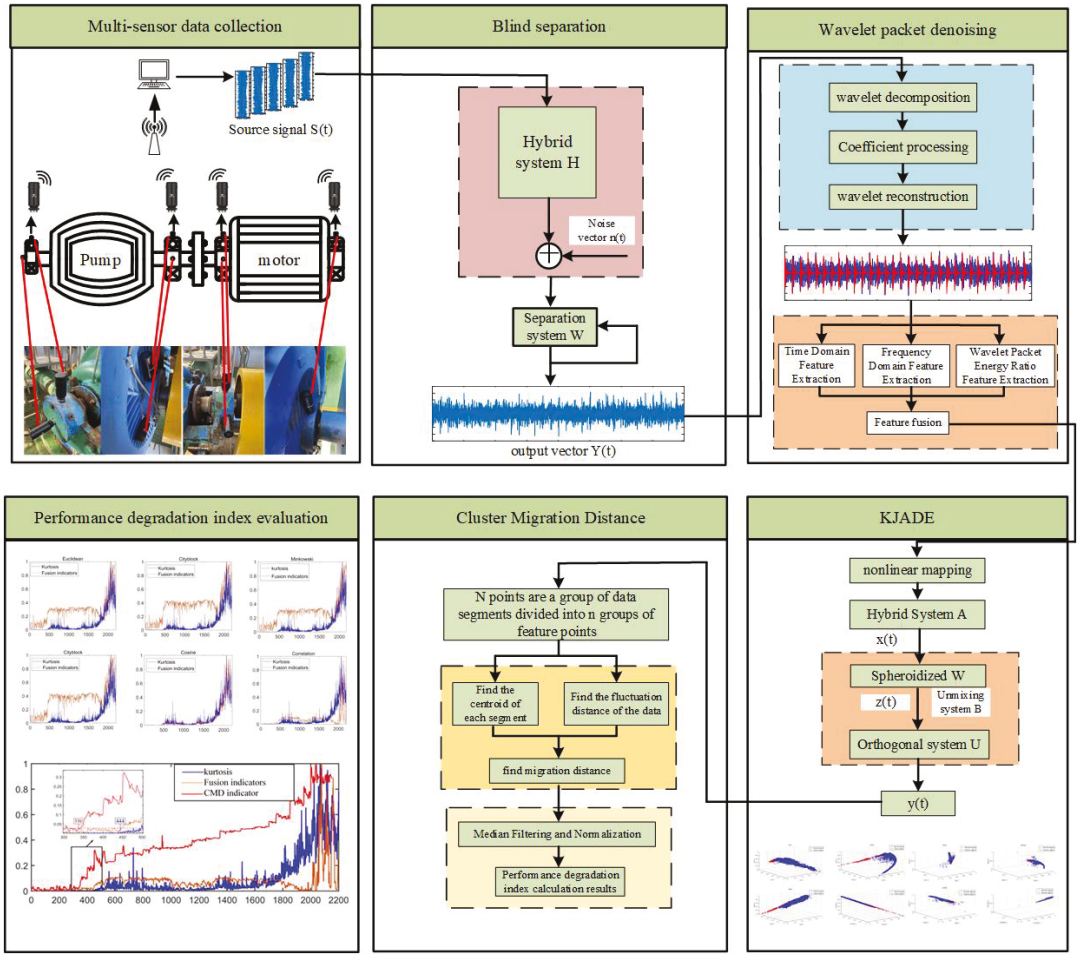


Figure 1. Overall flow chart.

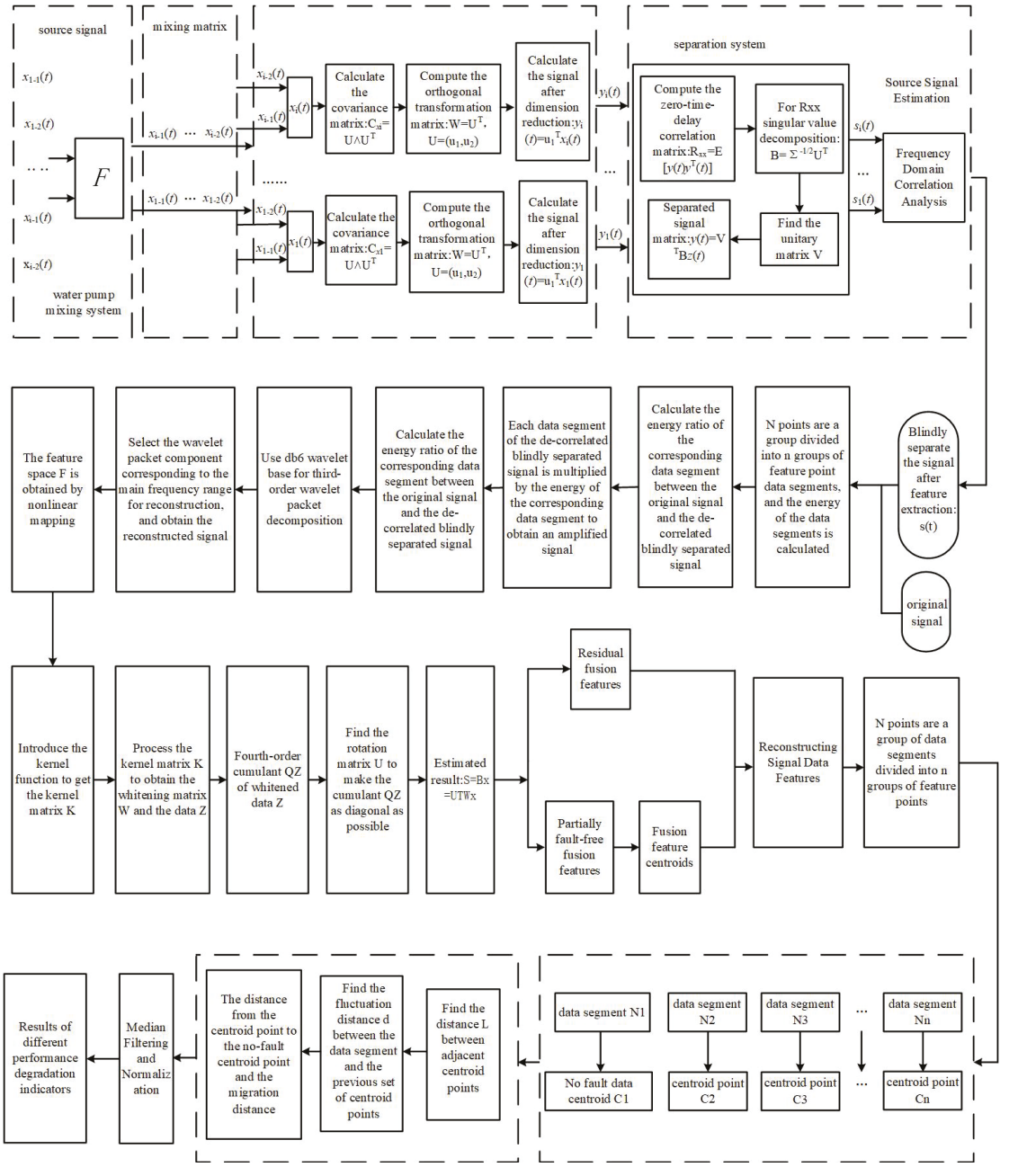


Figure 2. Overall algorithm flow chart.

Blind Separation of Decorrelation

In the previous link, the dimensionality reduction in the two-dimensional signals is measured by the horizontal and vertical displacement sensors of the same vibration measuring point. The signal matrix $\boldsymbol{y}(t)$ formed by each measuring point after dimension

reduction is used as the observed signal matrix for de-correlated blind separation, and feature extraction is achieved using de-correlated blind separation [18–20]. Assuming that the signals have statistical irrelevance, non-whiteness, or non-stationarity, and there is no more than one Gaussian signal in the signal, the eigenvalue decomposition of multiple non-zero time-delay correlation matrices is used to achieve blind signal separation. We then proceed as follows:

(1) The observed signal is pre-whitened. We calculate the zero-time delay correlation matrix R_{xx} of the observed signal matrix $\boldsymbol{y}(t)$ and perform singular value decomposition,

$$\boldsymbol{R_{xx}} = E\left[\boldsymbol{y}(t)\boldsymbol{y}^T(t)\right] = \boldsymbol{U}\sum\boldsymbol{U}^T, \boldsymbol{B} = \sum^{-1/2}\boldsymbol{U}^T \tag{4}$$

where Σ is a diagonal matrix consisting of singular values and then $z(t) = By(t)$.

(2) We find unitary matrices V and make the set of non-zero time-delay correlation matrices $\{R_{zz}(\tau_1), R_{zz}(\tau_2), \cdots, R_{zz}(\tau_k)\}$ joint diagonalization. The objective function is as follows:

$$f(\boldsymbol{V}) = \sum_{k=1}^{k} off\left(\boldsymbol{V}^H\boldsymbol{R}_{zz}(\tau_k)V\right) \tag{5}$$

Among them, k is the number of time delay correlation matrices, $off(\boldsymbol{M}) = \sum_{1<i\neq j<n}\left|M_{ij}\right|^2$.

(3) The separated signal matrix is:

$$\boldsymbol{s}(t) = \boldsymbol{V}^T\boldsymbol{z}(t) \tag{6}$$

The frequency domain correlation analysis is carried out between the separated signal and the monitored bearing signal, and the separated signal with a high-frequency domain correlation is taken as the monitoring bearing signal after blind separation.

2.2. Wavelet Packet Analysis to Achieve Shock Component Enhancement

After blind separation, the signal $s(t)$ is subjected to wavelet packet decomposition, and its decomposition tree is shown in Figure 3.

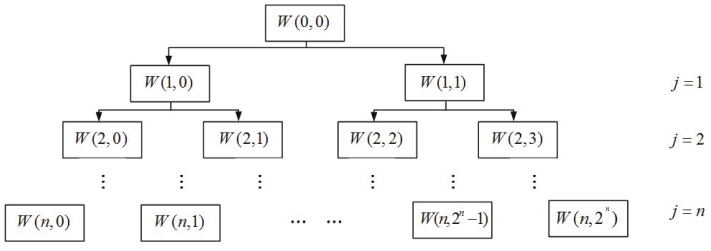


Figure 3. Wavelet decomposition tree.

If the sampling frequency of the bearing vibration signal is f_s, after the n-layer wavelet packet analysis, there are a total of 2^n nodes, and the frequency interval of each node segment is $\frac{fs}{2^{n+1}}$.

By analyzing the spectrum of the signal $s(t)$, the concentration area of the fault's frequency components is determined, and then the corresponding node is selected for reconstruction to obtain the signal $x(t)$.

2.3. Multivariate Feature Extraction in Time–Frequency Domain

After the vibration sensor collects the bearing signal, extracting the features reflecting the bearing state from these data is a key step in realizing fault diagnosis. The quality of the extracted features directly affects the recognition accuracy. Due to the complexity

of the bearing structure and the superposition and coupling of the vibration signals of various components, a single characteristic index, such as the root mean square value or peak value, cannot accurately reflect the current state of the bearing [21]. It can, however, comprehensively evaluate the running state of the bearing [22]. Therefore, the time-domain dimensionless index, the frequency-domain index, and the energy ratio index based on the wavelet packet sub-band are extracted as the original high-dimensional feature vectors.

2.3.1. Extraction Method of Time-Domain Features

The signal obtained by wavelet packet decomposition and reconstruction is $x(t)$, and n is the number of sampling points. The extracted signal feature set has dimensional indicators and non-dimensional indicators [23], as shown in Table 1.

Table 1. Time and frequency domain and time–frequency domain characteristic indicators.

Features	Expression	Features	Expression
Mean	$F1 = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$	Center frequency	$F13 = \sum_{i=1}^{N} f_i s_i / \sum_{j=1}^{N} s_j$
Root mean square value	$F2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2(t))}$	Frequency standard deviation	$F14 = \frac{1}{N}\sum_{j=1}^{N}(s_j - \frac{1}{N}\sum_{i=1}^{N} s_i)$
Square root magnitude	$F3 = \left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i(t)\rvert}\right]^2$	Root mean square frequency	$F15 = \sqrt{\sum_{i=1}^{N} f_i^2 s_i / \sum_{j=1}^{N} s_j}$
Absolute mean	$F4 = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i(t)\rvert$	First Band	$F17 = E_0/E$
Kurtosis	$F5 = \frac{1}{N}\sum_{i=1}^{N} x_i^4(t)$	Second Band	$F18 = E_1/E$
Variance	$F6 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - F1)^2$	Third Band	$F19 = E_2/E$
Waveform Indicator	$F7 = \frac{F2}{F4}$	Fourth Band	$F20 = E_3/E$
Peak indicator	$F8 = \frac{\max(\lvert x_i(t)\rvert)}{F2}$	Fifth Band	$F21 = E_4/E$
Impulse indicator	$F9 = \frac{\max(\lvert x_i(t)\rvert)}{F4}$	Sixth Band	$F22 = E_5/E$
Margin indicator	$F10 = \frac{\max(\lvert x_i(t)\rvert)}{F3}$	Seventh Band	$F23 = E_6/E$
Skewness Indicator	$F11 = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i^3(t)}{F2^3}$	Eighth Band	$F24 = E_7/E$
Kurtosis indicator	$F12 = \frac{F5}{F2^4}$		
Absolute mean	$F16 = \frac{1}{N}\sum_{i=1}^{N}[s_i(f_i - F13)^4]/(\frac{1}{N}\sum_{j=1}^{N}(s_j(f_i - F_{13})^2))^2$		

2.3.2. Extraction Method of Frequency Domain Features

When the bearing is in a healthy state, its vibration and acoustic signals are small during operation. When the bearing is damaged locally, a periodic impact signal will be generated, which will lead to the high-frequency vibration of the bearing itself. Therefore, the vibration signal of the bearing is very complex, and different types of bearing faults have different fault frequencies and impact laws. Therefore, the complex time domain signal can be transformed into a single harmonic component through Fourier transform for research, so as to obtain each harmonic component, such as the amplitude and phase information of the wave. Since the signals of different fault types do not have exactly the same frequency spectra, different frequency domain characteristic parameters are required for monitoring [24]. F13–F16, as listed in Table 1, are the frequency domain features required in this chapter, where f_i and s_i are the frequency and amplitude corresponding to the i-th spectral line of the reconstructed signal $x(t)$.

2.3.3. Extraction Method of Time–Frequency Domain Features

As a typical time–frequency domain analysis method, wavelet decomposition can perform multi-scale transformation of vibration signals. Since the fault information widely exists in different frequency components in the signal, a change in the frequency component often indicates that the state of the bearing has changed, so wavelet packet frequency band energy detection technology can be used to realize the original feature extraction of the bearing's operating state.

Third-order wavelet packet decomposition is performed on the reconstructed signal $x(t)$, and the energy corresponding to the node x_{3i} is E_i, while the corresponding discrete

point amplitude is h_{ik}, $i = 0, 1, 2, \ldots, 7; k = 1, 2, \ldots, n$, and n is the number of sampling points. Then, the energy of the i-th subband signal is:

$$E_i = \int |x_{3i}|^2 dt = \sum_{k=1}^{n} |h_{ik}|^2 \tag{7}$$

In order to more intuitively judge the change in energy and perform dimensionless processing on the extracted sub-band energy of wavelet packets, the energy ratio of each frequency band energy to the total energy E, as shown in Table 1, can be obtained as the time–frequency domain energy ratio original features, of which the total energy $E = \sum_{i=0}^{7} E_i$.

2.4. Advanced Feature Fusion Extraction Based on KJADE

In the previous section, time-domain indicators, frequency-domain indicators, and energy ratios based on wavelet packet sub-bands were extracted from the vibration signal, and formed a multi-domain feature set, avoiding the shortcoming of the insufficient evaluation capability of a single feature. However, there are redundant and conflicting problems among some features in multi-domain features, so it is necessary to extract feature quantities sensitive to bearing fault states. The bearing vibration signal often has nonlinear characteristics, and JADE belongs to the linear processing method, so it cannot effectively extract the characteristic quantities sensitive to the bearing fault state. In order to further apply JADE to bearing vibration signals with nonlinear behavior, the idea of the kernel function method is introduced to JADE, and the joint approximate diagonalization based on the kernel function eigenmatrix is obtained. Kernel Feature Matrix Joint Approximate Diagonalization (KJADE) is a novel feature fusion method. It not only has the characteristics of JADE, but also has better nonlinear processing capabilities than JADE. The core idea is to map the data $X \in R^{n \times m}$ to the high-dimensional feature space F through the nonlinear function Φ, and then use the JADE algorithm to transform the nonlinear separable problem into a linearly separable problem in F. The mapping process is shown in Figure 4. Assuming that the sample space is $X = \{x_1, x_2, \ldots, x_n\}$, where x_i is the input vector of the i-th dimension of the sample space, which contains n data points, after mapping, the feature space is $F = \{\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_M)\}$.

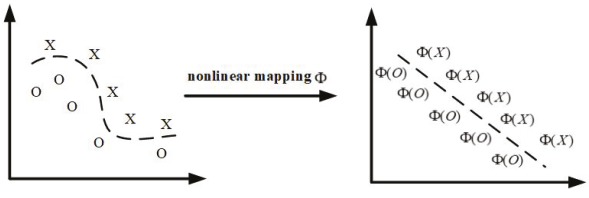


Figure 4. Nonlinear mapping.

For the same purpose as JADE, KJADE also needs to find unmixed matrix B to obtain the optimal matrix, namely:

$$y = B^F \Phi(x) = U^T W \Phi(x) \tag{8}$$

Furthermore, the covariance matrix of the mapped F can be obtained as:

$$R_F = \frac{1}{N} \Phi(x_i) \Phi(x_i)^T = \frac{1}{N} F F^T \tag{9}$$

R_F and its eigenvalues and eigenvectors cannot be obtained due to the "curse of dimensionality" caused by the excessive dimension of the feature space. The idea of kernel

function is introduced here, and the complex and time-consuming inner product calculation is converted into a kernel function, and an $N \times N$ kernel matrix K is obtained:

$$K_{ij} = \Phi(x_i)\Phi(x_j) = k(x_i, x_j) \tag{10}$$

Among them, K_{ij} must meet the Mercer condition, that is, $K = FF^T$. The following are commonly used kernel functions:

Gaussian kernel function:

$$k(x_i, x_j) = \exp(-\frac{x_i - x_j^2}{2\sigma^2}) \tag{11}$$

Polynomial Kernel Function:

$$k(x_i, x_j) = (ax_i^T x_j + c)^d \tag{12}$$

Sigmoid Kernel function:

$$k(x_i, x_j) = \tan h(ax_i^T x_j + c) \tag{13}$$

Since the Gaussian kernel function achieves better results in solving practical problems [25], the Gaussian kernel function is used in this study to replace the inner product operation, where σ represents the width parameter of the function. Then, the eigendecomposition of the fourth-order cumulant matrix of the extracted kernel matrix is produced, and the nonlinear feature fea_kjade(x_i,y_i,z_i) hidden in the observation signal is obtained.

2.5. Calculation of CMD Performance Degradation Index

When the bearing fails, it has good class separability from the characteristic distribution of the healthy bearing. With the continuous operation of the bearing, the failure degree is expanded, and the characteristic difference between the monitored bearing and the healthy bearing becomes larger. Therefore, two types of models can be constructed to evaluate the difference between monitoring signals and health signals. In the classification of samples, the intra-class distance and the inter-class distance have been applied in the class separability measurement between different bearing fault types.

The two types of models constructed are shown in Figure 5. In the extracted nonlinear feature fea_kjade(x_i,y_i,z_i), we extract the feature segment of the bearing in normal operation and record the feature set extracted in the healthy state of the bearing as X_o, the feature set extracted by the bearing at time t is Y_t, and then the two types of models formed are $Z_t = [X_o, Y_t]$, where $X_o, Y_t = (x_1, x_2, \ldots, x_i, \ldots, x_n)$, n is the number of samples, $x_i \in R^D$ (D is the feature dimension). Then, the inter-class scatter matrix is:

$$S_b = \sum_{i=1}^{c} P_i m_i - m^2 \tag{14}$$

The intra-class scatter matrix is:

$$S_w = \sum_{i=1}^{c} P_i \frac{1}{n_i} \sum_{k=1}^{n_i} x_k^i - m_i^2 \tag{15}$$

In:

$$\begin{cases} P_1 = n_i / \sum_{j=1}^{c} n_j \\ m_i = 1/n_i \sum_{k=1}^{n_i} x_k^i \\ m = \sum_{i=1}^{c} p_i m_i \end{cases} \quad 1 \le i, j \le C \tag{16}$$

C is the number of classes, where $C = 2$, and m_i and m are the feature mean of class i and the entire sample, respectively. The inter-class scatter matrices S_b and S_w represent the degree of aggregation between different classes and between the same class, respectively.

To describe the feature part more comprehensively, the intra-class and inter-class distance *SS* is used as a measure of class separability, and its expression is shown in Equation (17).

$$SS = trace(S_b / S_w). \tag{17}$$

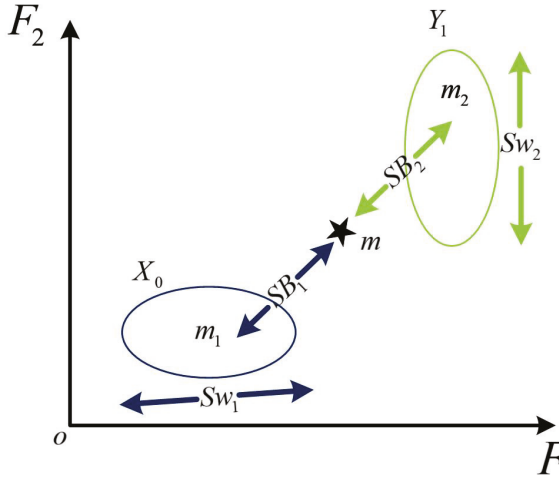


Figure 5. Schematic diagram of the two types of models.

2.6. Performance Degradation Indicator

2.6.1. Selection of Distance Index Calculation

Because clustering does not know any sample labels, the samples are divided into different classes through the internal relationship and different calculation methods are used, and the clustering will also obtain different results. The following are commonly used similarity calculation methods. Therefore, when calculating small sample clustering, it is necessary to calculate the relevant distance. When calculating the distance between the feature set Y_t extracted at time t and the centroid of the feature set X0 extracted under the healthy sample, we use the following distance formulas as a reference for comparison:

1. Euclidean Distance

According to the calculation, this distance can be considered to be the L2 norm. Two points a(x_1,x_2 … x_n) and b(y_1,y_2 … y_n) are in the n-dimensional space:

$$d_{ab} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{18}$$

2. Manhattan Distance

According to the calculation, this distance can be considered to be the L1 norm. Two points a(x_1,x_2 … x_n) and b(y_1,y_2 … y_n) are in the n-dimensional space:

$$d_{ab} = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| \tag{19}$$

3. Chebyshev Distance

It is simply considered the maximum value of the coordinate difference of each coordinate. Two points a(x_1,x_2 … x_n) and b(y_1,y_2 … y_n) are in the n-dimensional space:

$$d_{ab} = max(|x_1 - y_1|, |x_2 - y_2|, \ldots |x_n - y_n|) \tag{20}$$

4. Minkowski Distance

The Minkowski distance between two n-dimensional variables a(x_{11},x_{12}, … ,x_{1n}) and b(x_{21},x_{22}, … ,x_{2n}) is defined as:

$$d_{12} = \sqrt[p]{\sum_{k=1}^{n} |x_{1k} - x_{2k}|^p} \tag{21}$$

where p is a variable parameter. Min's distance defines a set of distance formulas, including the Euclidean distance, Manhattan distance, and Chebyshev distance.

5. Cosine Distance

The cosine similarity derivation formula is as follows:

$$\cos(\theta) = \frac{a^2 + b^2 - c^2}{2ab} \tag{22}$$

Two points a(x_1,x_2 … x_n) and b(y_1,y_2 … y_n) are in the n-dimensional space

$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} {x_{1k}}^2} \sqrt{\sum_{k=1}^{n} {x_{2k}}^2}} \tag{23}$$

6. Correlation Distance

The correlation coefficient is a way to assess the degree of correlation between random variables X and Y.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X - EX)(Y - EY)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{24}$$

Correlation distance:

$$Dxy = 1 - \rho_{XY} \tag{25}$$

2.6.2. Dynamic Centroid

When the bearing is running, its characteristic signal data migrate and change. In order to more clearly observe the migration and change trend of the fusion features, we propose the concept of the dynamic centroid. When finding the centroid of a single data cluster, it is often best to find a fixed number of K centroid points for the entire data cluster and iteratively find a partition scheme for K clusters, minimizing the loss corresponding to the clustering result function, where the loss function can be defined as the sum of squared errors between the sample points in each cluster and their center points:

$$J(c, \mu) = \sum_{i=1}^{M} {x_i - \mu_{ci}}^2 \tag{26}$$

where x_i represents the i th sample, c_i is the cluster to which x_i belongs, μ_{ci} represents the center point corresponding to the cluster, and M is the total number of samples.

As shown in Figure 6, first, we randomly select a sample point as the initial centroid, and then calculate the similarity between each sample point and the centroid. We classify each sample point into its most similar category, then recalculate the centroid point (i.e., the class center) of each class again, repeat this process until the centroid point no longer changes, and finally, the class and the class centroid can be obtained. We then divide the entire data cluster into a given number of K classes, but this class, divided according to data characteristics, may not have the trend we want, the size of each class may not be the same, and its data points may not be the same or continuous. We refer to this idea when calculating the migration trend of the feature distance, divide the data into n segments, find the centroid point for each segment, divide the entire data segment into n sample clusters, and in each sample cluster, the fluctuation distance between the data point and the centroid is calculated. The sum of the movement distance of the centroid and the fluctuation distance is the migration distance we want.

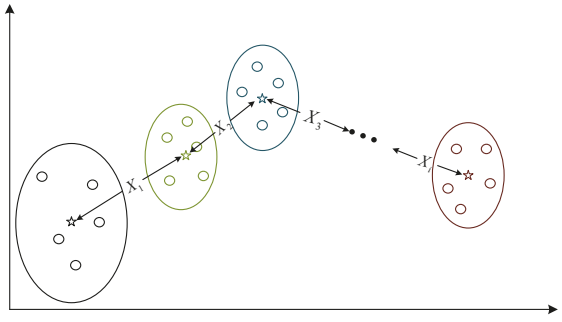


Figure 6. Schematic diagram of dynamic centroid algorithm.

Therefore, here we divide the nonlinear fusion feature fea_kjade(x_i,y_i,z_i), extracted from the feature fusion in the previous step, into n sample clusters, find the centroid C_0 of the fault-free sample cluster, find the centroid point for each sample cluster separately C_i, find the migration distance X_i between the two centroid points, and find the fluctuation distance d_i between the sample cluster and the previous centroid point C_{i-1}. At this time, the required performance degradation evaluation index ddd_1 can be obtained by adding these two distances.

The basic idea is to select the early fault-free samples initially judged in the KJADE fusion feature index as the fault-free sample clusters, and then divide the remaining samples into sample clusters according to the predetermined number of samples, calculate the centroids of the sample clusters, and specify each sample. The distance to its corresponding sample cluster centroid plus the migration distance from the sample cluster centroid point to the non-faulty sample cluster centroid point is the dynamic centroid distance of the sample. The pseudocode of the dynamic centroid algorithm is shown in Algorithm 1.

Algorithm 1 Algorithm name: Dynamic centroid algorithm

Input: KJADE feature feakjade after input wavelet packet decomposition and reconstruction.
1: n; Number of sample clusters.
2: N;The number of samples contained in each sample cluster.
3: [idx, C_0] = $kmeans$(fea_kjade(:, 1:350)′, 1); Find the coordinates of the no-fault centroid point.
4: **for** i = 1, 2, . . . n **do**
5: [idx, C_1] = $kmeans$(fea_kjade(:, 351 + N ∗ (i − 1):350 + N ∗ i)′, 1); Find the coordinates of the segmental centroid point.
6: **end for**
7: d(1, :) = $pdist2$(C_0, C(1, :), ′$euclidean$′); Find the distance between the first two centroids
8: **for** i = 1, 2, . . . n − 1 **do**
9: d(i + 1, :) = $pdist2$(C(i, :), C(i + 1, :), ′$euclidean$′); Find the coordinates of the two centroid points.
10: **end for**
11: d_3(1, 1:350) = $pdist2$(C0, $feakjade$(:, 1:350)′, ′$euclidean$′); Find the migration distance between the fault-free data segment and the initial centroid point.
12: **for** i = 2, 3, . . . n − 1 **do**
13: d_1 = $pdist2$(C(i, :), $feakjade$(:, 351 + N ∗ (i − 2):350 + N ∗ (i − 1))′, ′$euclidean$′); Find the fluctuation distance of the data.
14: **for** j = 1, 2, . . . i **do**
15: d_2 = sum(d(1:j, :)); Calculate the sum of the distance of the centroid.
16: **end for**
17: d_3(1, 351 + N ∗ (i − 2):350 + N ∗ (i − 1)) = d1 + d2; Calculate the total distance.
18: **end for**
19: v1 = $medf ilt$1(d_3, 5); Median filter
20: ddd1 = $mapminmax$(v1, 0, 1); Normalized
Output: Performance degradation indicator graph ddd1

3. Results and Discussion

This paper takes the vibration signals collected by each measuring point of a pump from 11 September to 24 December 2021 as the analysis object. The physical diagram of the equipment structure of the pump set and the installation of the vibration sensor is shown in Figure 7. Vibration sensors in horizontal and vertical directions are placed on the free end and the driving end of the water pump and drive motor. The measuring point data sets are 1H, 2H, 2V, 3H, 3V, 4H, and 4A. Each half of each measuring point collects group data, eliminates the useless data due to problems such as shutdown and acquisition terminal failure during this period, and obtains 3684 groups of vibration signals during this period to complete the data preprocessing. The technical parameters related to the water pump are shown in Table 2, and the overall structure and physical diagram are shown in Figure 7.

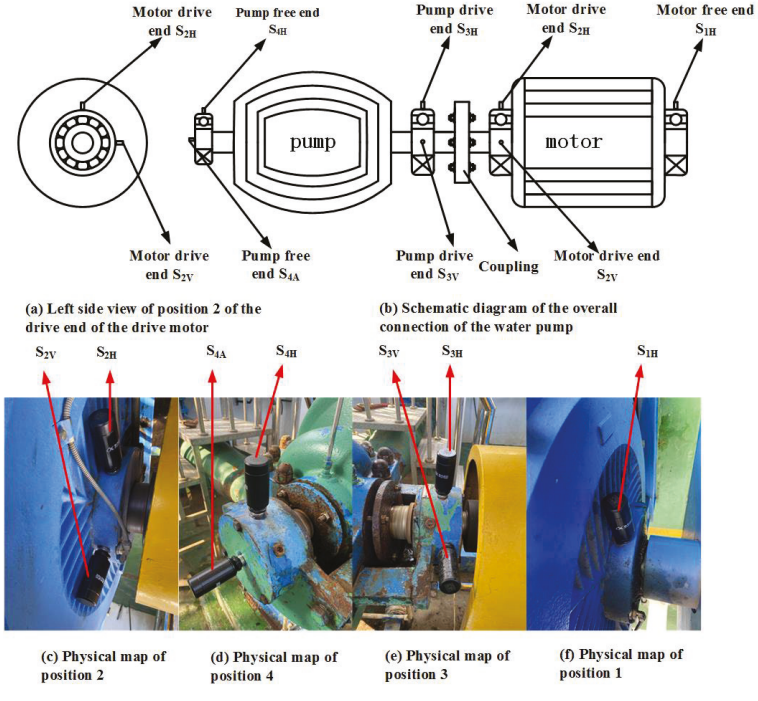


Figure 7. Water pump set equipment structure and vibration sensor installation physical diagram. (**a**) Left side view of position 2 of the drive end of the drive motor; (**b**) Schematic diagram of the overall connection of the water pump; (**c**) Physical map of position 2; (**d**) Physical map of position 4; (**e**) Physical map of position 3; (**f**) Physical map of position 1.

Table 2. Pump-related technical parameters.

Items	Parameters	Items	Parameters
power	220 Kw	COS ∅	0.85
frequency	50 HZ	work schedule	S1
Phase	3	Insulation class	155
Voltage	6000 V	Rotating speed	1490 r/min
current	26.8 A	cooling method	IC611
Rotating speed	1480 r/min	Diameter of impeller	448 mm
Equipped with power	220 KW	flow	720 m^3/h

Table 2. *Cont.*

Items	Parameters	Items	Parameters
Rotating speed	1490 RPM	frequency conversion	24.8 HZ
BPFI	121.6 HZ	BPFO	76.9 HZ
BSF	51.6 HZ	FTFO	9.43 HZ

3.1. Raw Signal Analysis

Figure 8 is the original signal time-domain diagram and partially enlarged diagram. From the partial magnification of the time domain signal, we can see the minor fault segment, that is, the data segment whose amplitude has not yet increased but the periodic shock component begins to appear. The coordinates of the corresponding sample points are 5,521,700 to 6,856,630. The envelope spectrum of the minor fault signal segment is shown in Figure 9. From the figure, one can observe the obvious cage fault frequency and rotation frequency and its frequency multiplication component. Regarding the sample of the mid-term fault section, the coordinates of the corresponding sample point are 24,000,000 to 25,000,000, and the envelope spectrum is drawn as shown in Figure 10, in which the outer raceway failure frequency, inner raceway failure frequency, rolling element failure frequency, and cage failure can be observed. This shows that in the severe fault section, the bearing fault evolved from the cage fault to a composite fault composed of the outer raceway fault, the inner raceway fault, the rolling element fault, and the cage fault. Regarding the samples of the later fault section, the coordinates of the corresponding sample points are 50,000,000 to 51,000,000, the main frequency component in the envelope spectrum is the rotation frequency, and there are sidebands around the rotation frequency, indicating that modulation occurs in the severe fault section.

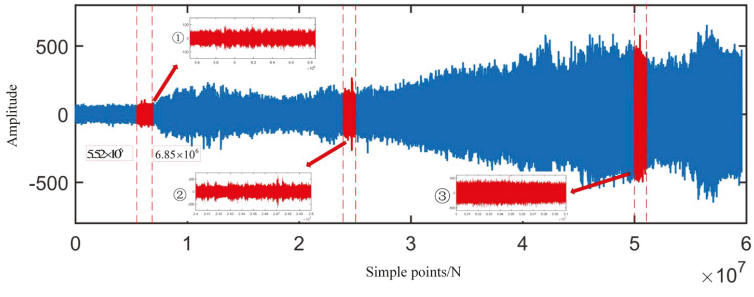


Figure 8. Time domain diagram of the original signal.

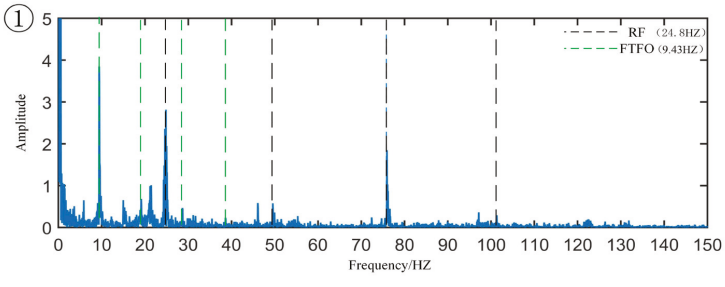


Figure 9. Minor fault signal envelope spectrum.

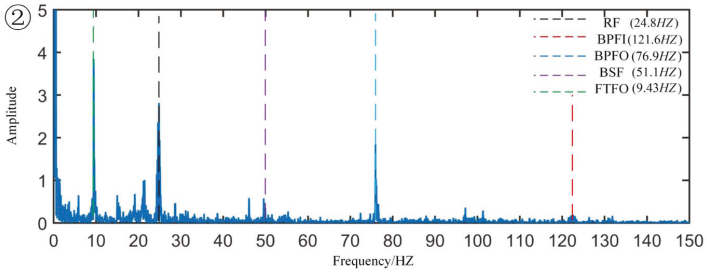


Figure 10. Mid-term fault signal envelope spectrum.

For the minor fault segment, according to the number of sample points n = 16,384, the kurtosis value of the segment is calculated to draw the kurtosis curve of the original signal, as shown in Figure 11. The earliest increase in kurtosis occurs at sample segment 437, and the corresponding sample point coordinate is 7,159,808, indicating that a single kurtosis index has limitations in the early fault diagnosis of bearings and cannot accurately identify minor faults of bearings.

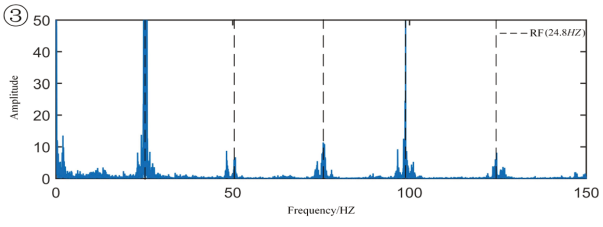


Figure 11. Later fault signal envelope spectrum.

3.2. Analysis of Blind Separation and Wavelet Packet Reconstruction Processing Results

In order to eliminate the interference of other vibration sources in the pump device to the signal acquisition under actual working conditions, the source signal is initially processed by the blind separation processing method, and wavelet packet decomposition is performed on the collected bearing vibration signal after blind separation. The decomposition structure is shown in Figure 12, where W(j, m) represents the node m of the layer j, each node represents the decomposition coefficient of the original signal $s(t)$ on the scale j for the wavelet packet function, m represents the frequency band, and each node is associated with the corresponding frequency band match. In the figure, W(0, 0) represents the original signal, and other nodes such as W(2, 0) represent the 0th node coefficient of the second layer, and then each subband signal is extracted by reconstruction.

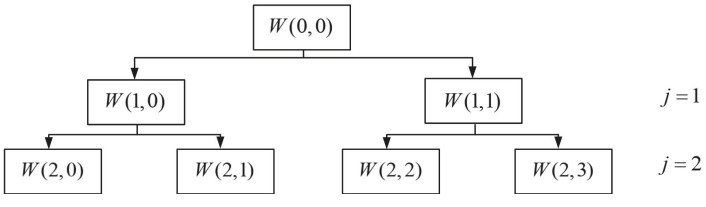


Figure 12. Wavelet decomposition tree.

When the number of nodes m is an even number, it represents the low-frequency component signal decomposed by the low-pass filter coefficient h(k). On the contrary, when

m is an odd number, it represents the high-frequency component signal decomposed by the high-pass filter coefficient g(k).

The sampling frequency of the bearing vibration signal is 12,800 Hz. The wavelet packet is decomposed into two layers, while the third layer has 2 ^ 2 = 4 frequency segments, and the frequency range of each frequency segment is 6400/4 = 1600 Hz. The specific range is shown in Table 3.

Table 3. Frequency range.

Node	W(2,0)	W(2,1)	W(2,2)	W(2,3)
Frequency range/HZ	0~1600	1601~3200	3201~4800	4801~6400

It can be seen from the table that the frequency bands corresponding to different nodes are different. Therefore, we analyze the spectrum of the signal after blind separation, create its spectrogram, observe the concentrated segment of the fault frequency, and select the wavelet node for the next step.

The bearing signal is a typical irregular signal. In order to grasp the time–frequency characteristics of each frequency band, the dbN wavelet can be used for multi-layer decomposition, and the time–frequency analysis of the signal is carried out in different frequency ranges. Because the db6 wavelet is a tightly supported orthogonal real wavelet, with good regularity and a large vanishing moment, it is used as the wavelet basis for wavelet packet decomposition. The second-order wavelet packet decomposition is performed on the signal after blind separation, and the db6 wavelet is selected as the wavelet base. Figure 13 is the spectrogram of the signal after blind separation processing. Through spectrogram analysis, the main frequency component of the signal is approximately 840 HZ. Therefore, node 4 (frequency band 1–1600 HZ) in the tree node is selected for reconstruction.

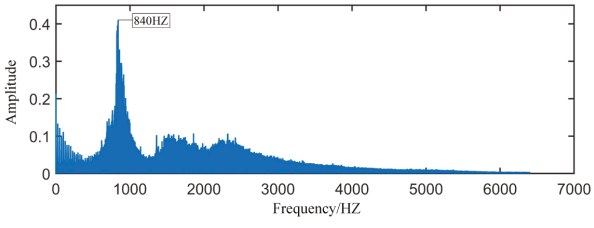
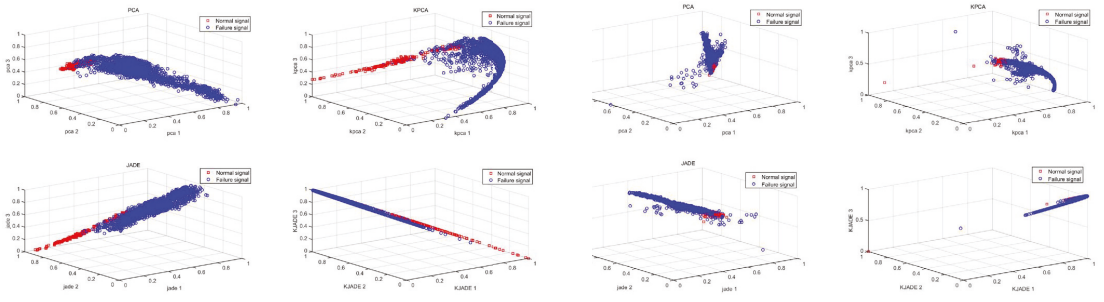


Figure 13. Spectrum of the signal after blind separation.

To better verify the advantages of reconstructed signals, PCA, KPCA, JADE, and KJADE methods are used to fuse the original data and the reconstructed data, and the processed feature points are normalized and the first four are preliminarily selected. The normal working data of the bearing is used as a no-fault signal. Figure 14 shows the feature point clustering effect after feature fusion of source data and reconstructed data. Table 4 shows the intra-class distance of KJADE and JADE. By calculating the intra-class and inter-class distances between different feature points, it is found that the intra-class inter-class distances of the non-faulty samples and early fault samples in the reconstructed signal are significantly lower than the original signal, indicating that after blind separation and wavelet packet reconstruction, the clustering effect of feature points obtained by feature fusion of the four methods, PCA, KPCA, JADE, and KJADE, has been significantly improved.



(**a**) Raw data (**b**) Reconstructed data

Figure 14. Feature point clustering.

Table 4. Intra-class distance.

	Intra-Class Distance of KJADE	Intra-Class Distance of JADE
Raw data	0.7606	1.4764
reconstruct data	0.0776	0.8045
	Intra-Class Distance of KPCA	**Intra-Class Distance of PCA**
Raw data	1.043	0.9804
reconstruct data	0.8771	0.7566

3.3. Fusion Feature Distance Metrics

Considering that the three-dimensional coordinates of the signal feature points can only reflect the distribution and clustering effect of the feature points in the three-dimensional space, it cannot intuitively reflect the time node when the bearing starts to fail, and the trend of the subsequent bearing failure severity changes with time. The normal working data of the bearing in the first four days is used as the no-fault sample cluster, the centroid coordinates of the no-fault sample cluster are obtained, the distance between each subsequent feature point and the centroid of the no-fault sample cluster is calculated, and the distance index is used as the judgment bearing. A new indicator of the failure level is obtained.

To demonstrate the superiority of KJADE in the distance trend after the fusion of bearing fault features, feature distance calculation processing was performed on the data after feature fusion using PCA, KPCA, JADE, and KJADE methods. In terms of calculation, six different distance calculation methods, such as the Euclidean distance, Manhattan distance, and correlation distance, are used for verification.

Figure 15 is a trend diagram of the fusion feature distance indicators of the four feature fusion indicators PCA, KPCA, JADE, and KJADE under different distance calculation methods. The PCA fusion feature distance index has a large amplitude at the non-fault sample cluster, which cannot accurately judge the fault occurrence time. The angle cosine distance index and correlation distance index of the KPCA fusion feature have low amplitudes at the early non-fault sample clusters, which are in line with the actual working conditions, but the fault occurrence point is consistent with the kurtosis index, and early faults cannot be diagnosed in advance. There are also cases where the amplitudes of the remaining distance indicators are too large at the fault-free samples. The overall trend of JADE fusion features fluctuates greatly, and the amplitude is large at the fault-free sample cluster, which will interfere with the judgment of bearing faults. The Chebyshev distance index, the included angle cosine distance index, the Euclidean distance index, the Manhattan distance index, and the Minkowski distance index of the KJADE fusion feature also have the same problems as the above three other feature fusion methods. The amplitude at the cluster is abnormal, and the early warning of the early bearing failure cannot be realized.

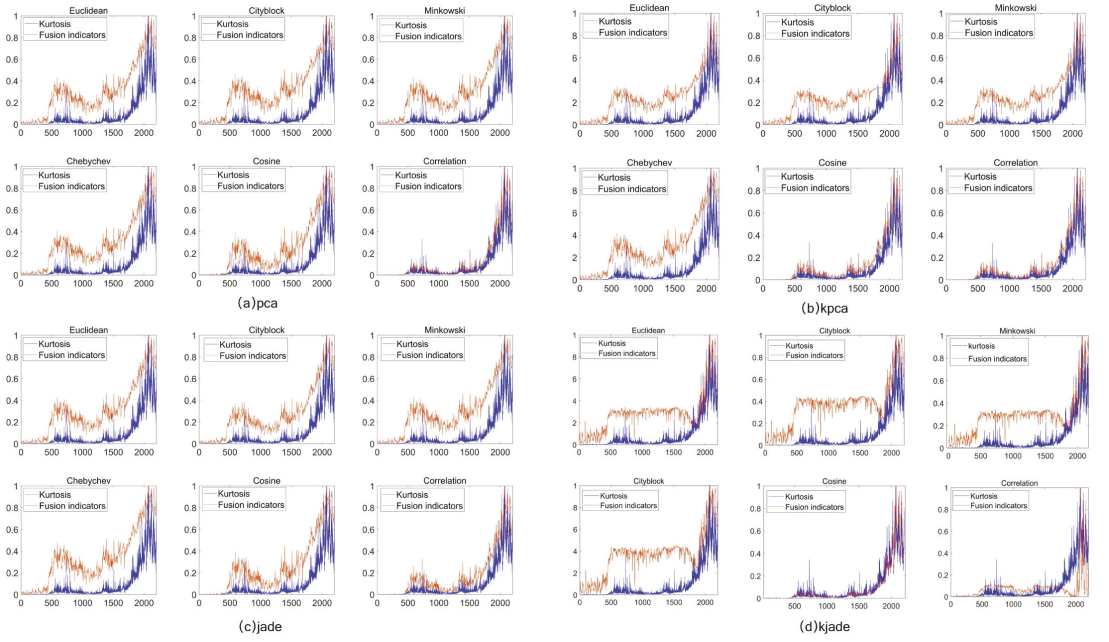


Figure 15. The effect of the comparison between the distance index and the cliff index without the distance calculation method. (**a**) PCA; (**b**) KPCA; (**c**) JADE; (**d**) KJADE.

The relevant distance index of KJADE effectively realizes the early warning of the early failure of the faulty bearing. Figure 16 shows a partially enlarged view of the correlation distance index of the KJADE fusion feature. The correlation distance index of the KJADE fusion feature rises at sample segment 336, and the corresponding sample coordinate is 5,505,024, which is the same as the starting sample coordinate of the minor fault segment in Section 3.1. 5,521,700 is basically the same, indicating that the correlation distance index of the KJADE fusion feature realizes the early diagnosis of minor bearing faults.

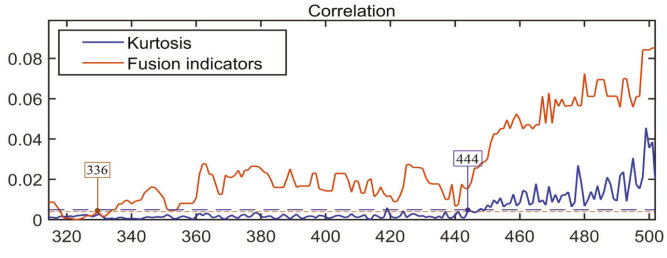


Figure 16. Local enlargement of the correlation distance index of the KJADE fusion feature.

3.4. Cluster Migration Distance

In Section 3.2, after the early failure occurred, the kurtosis index and the KJADE fusion index showed a trend of first rising and then decreasing, and this anomalous phenomenon did not match the actual situation. In order to more accurately describe the trend of bearing failure degree with time, based on KJADE feature fusion, this paper proposes a novel performance degradation metric extraction method called Cluster Migration Distance (CMD). On the premise that the sample points corresponding to minor faults are determined

in Section 3.2, all non-fault sample points are selected to obtain their centroids, divide the remaining sample points into n sample clusters according to the specified sample cluster capacity N, calculate the centroid of each sample cluster, and calculate the distance $X_i(i = 1, 2, 3, \ldots, N)$. In the subsequent calculation of the distance index of the sample points, the dynamic centroid method is used, that is, the distance $x_j(j = 1, 2, 3 \ldots)$ from each sample point to the centroid of the sample cluster where the sample point is located. The migration distance of the centroid point of the sample cluster is used as the dynamic centroid distance v_j of the sample point. The specific calculation formula is as follows:

$$v_j = \sum_1^{[j/n]+1} X_i + x_j \tag{27}$$

Considering the calculation accuracy and the total number of samples, this paper takes the sample cluster capacity n = 50 and the number of sample clusters as 37, calculates the dynamic centroid distance of each sample point, and draws its change trend as shown in Figure 17. Compared with the KJADE integrated index and the kurtosis index, the dynamic centroid distance index has a larger minor fault amplitude, which allows it to make a more accurate judgment on the minor fault of the bearing, and the subsequent amplitude shows an upward trend, which is more in line with actual engineering conditions.

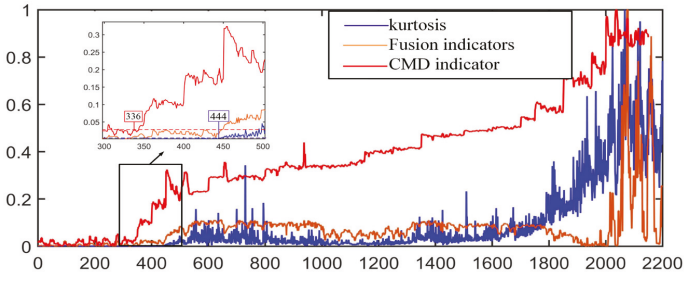


Figure 17. Dynamic center-of-mass metrics.

In the early stages of equipment operation, the equipment operates normally, and various vibration indicators fluctuate normally. The fluctuations originate from changes in the flow and load of the pump. The bearing at the drive end of the pump has early normal wear characteristics, but there is no deterioration trend. With the operation of the pump, the corresponding fault information of the pump vibration data, such as kurtosis and other indicators, change significantly compared to the normal operation stage. At this time, the bearing will result in early failure. As the faulty bearing continues to work, various indicators fluctuate at high and low points. Intensified through the analysis of the data, the time-domain waveform can be seen in the time-domain waveform with insignificant swarm impact characteristics and the appearance of impact instability, indicating that the bearing is worn at this time, but the damaged surface of the rolling element is small, and the load area occasionally makes contact during the rotation process. When the inner and outer raceways are in contact, the vibration acceleration index rises, and when it is not in contact, the indicators decline.

As the bearing continues to work, the high-frequency vibration kurtosis density and impact energy ratio of the pump drive end change again compared to the previous ones. All kinds of indicators show a significant upward trend, and the vibration model shows new changes. The impact of the quasi-rotation frequency interval can be seen in the high-frequency vibration acceleration waveform, which appears to be stable, and the acceleration envelope spectrum is dominated by the rotation frequency and the fault frequency of the bearing inner ring, indicating that the main damage of the bearing at this stage is on the bearing inner ring. At the same time, the rising indicators show that the bearing damage surface is gradually expanding. The rise in vibration energy mainly comes from the bearing

fault frequencies and sidebands in the frequency spectrum. The vibration velocity trend of the driving end and the free end of the pump shows that the effective value of the vibration velocity at the free end also shows a slow upward trend with time. This feature indicates that the vibration generated by the damage of the bearing at the driving end affects the entire shaft system of the pump.

It can be seen from Figure 17 that when the water pump is in a fault-free state, the distance between the fusion feature and the non-fault data center of mass is small. When the bearing is in a fault state, there will be a sudden change in the distance between the fusion coordinates and the non-fault center of mass, which can be determined at the moment of the sudden change, when the bearing failed. Among different fusion methods, we can see that compared with other feature fusion methods, the KJADE method is better than other feature fusion methods. When observing the trend of the feature distance, its change trend is more obvious. After the mutation, it can be clearly reflected by the relevant characteristic trend, so the state of the bearing during operation can be sensed in advance through the corresponding characteristic distance change. When the characteristic distance begins to grow larger or even begins to mutate, we can judge accordingly whether the bearing has failed.

When the kurtosis index is simply used, and when the number of sample points is 444, the kurtosis index begins to show a clear upward trend. At this time, the corresponding bearing failure stage is the mid-term failure of the bearing, and the early failure of the bearing cannot be accurately identified. At this time, the water pump bearing is running in a faulty state. The CMD index we proposed began to show an upward trend when the number of sampling points was 336, which was 20 h earlier than the simple kurtosis index. Early warning can be achieved, and economic losses caused by potential safety hazards can be prevented in time.

4. Conclusions

In practical engineering applications, the pump bearing signal is seriously disturbed by noise, and the fault evolution is complex and changeable. Traditional performance degradation indicators cannot describe the degradation trend in a timely and accurate manner. To solve this problem, this work proposes a novel CMD index extraction method. This method can effectively eliminate the interference of the coaxial vibration source and can accurately describe the degradation trend of the bearing. The analysis results of an actual engineering case show that its early warning ability and index monotonicity are better than the traditional kurtosis index.

However, there are still some future challenges: (1) Better denoising of the pump bearing signal and extracting effective degradation indicators to achieve earlier warning and more accurate fault evolution process characterization, which is still a future challenge; (2) improving the computational efficiency of the method to achieve real-time monitoring, which requires further research; (3) in the process of CMD calculation, some parameters need to be manually selected, and the method of realizing parameters' self-optimization deserves further study; and lastly, (4) the automatic selection and fusion of features is also worthy of further research to improve the clustering of features.

Author Contributions: Conceptualization, Z.Z. (Zhongping Zhai); Methodology, Z.Z. (Zhongping Zhai) and F.L.; Supervision, Z.F.; Validation, Z.Z. (Zihao Zhu), Y.X. and X.Z.; Writing–original draft, Z.Z. (Zhongping Zhai), Z.Z. (Zihao Zhu), Y.X. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by National Natural Science Foundation of China grant number [51875001].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Peng, G.; Guo, L.; Zhang, G.; Chen, X. Time-frequency analysis of vibration signal of water pump bearing. In *New Century, New Opportunities, New Challenges—Knowledge Innovation and High-tech Industry Development (Volume 2) Proceedings of the 2001 Academic Annual Conference of China Association for Science and Technology, Changchun, China, 13 September 2001*; China Science and Technology Press: Changchun, China, 2001; p. 331.
2. Omri, F.; Choura, O.; Hadj Taieb, L.; Elaoud, S. Prediction of Bearing Fault Effect on the Hydraulic Performances of a Centrifugal Water Pump. *J. Vib. Eng.Technol.* **2022**, 1–11. [CrossRef]
3. Chen, L.; Wei, L.; Wang, Y.; Wang, J.; Li, W. Monitoring and Predictive Maintenance of Centrifugal Pumps Based on Smart Sensors. *Sensors* **2022**, *22*, 2106. [CrossRef] [PubMed]
4. Dutta, N.; Kaliannan, P.; Subramaniam, U. Bearing Fault Detection for Water Pumping System Using Artificial Neural Network. In Proceedings of the 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 1–3 March 2022; pp. 1–6.
5. Tang, J.; Wang, E.; Zhu, J.; Tan, W. Research on fault diagnosis of automobile water pump bearing based on frequency domain feature and support vector machine. *Mach. Tool Hydraul.* **2018**, *46*, 163–167+55.
6. Li, T.; Duan, L.; Zhang, D.; Zhao, S.; Huang, H.; Bi, C.; Yuan, Z. Application of adaptive convolutional neural network in fault diagnosis of rotating machinery. *Vib. Shock.* **2020**, *39*, 275–282+88.
7. Xu, Y.; Zhang, F.; He, Z. Independent component analysis and its application in fault diagnosis. *Vib. Shock.* **2004**, *23*, 104–107.
8. Xu, T.; Pei, X.; Xu, T.; Lang, X. Application of wavelet neural network in fault diagnosis of double suction centrifugal pump bearing. *Coal Min.Mach.* **2011**, *32*, 254–257.
9. Han, H.; Cheng, D.; Xu, H. Fault diagnosis of feed water pump bearing based on stacked noise reduction and self-encoding. *Mechatron. Eng. Technol.* **2021**, *50*, 254–258.
10. Bie, Z.; Li, Q. Fault diagnosis of circulating water pump bearing based on peak energy. *Gen. Mach.* **2011**, 47–49.
11. Lyu, P.; Zhang, K.; Yu, W.; Wang, B.; Liu, C. A novel RSG-based intelligent bearing fault diagnosis method for motors in high-noise industrial environment. *Adv. Eng. Inform.* **2022**, *52*, 101564. [CrossRef]
12. Buchaiah, S.; Shakya, P. Bearing fault diagnosis and prognosis using data fusion based feature extraction and feature selection. *Measurement* **2022**, *188*, 110506. [CrossRef]
13. Sun, R.B.; Du, F.P.; Yang, L.H.; Ma, M.; Yang, Z.B.; Chen, X.F. Feature-guided regularization parameter selection in sparse de-noising for fault diagnosis. *Mech. Syst. Signal Process.* **2022**, *179*, 109373. [CrossRef]
14. Yang, C.; Ma, J.; Wang, X.; Li, X.; Li, Z.; Luo, T. A novel based-performance degradation indicator RUL prediction model and its application in rolling bearing. *ISA Trans.* **2022**, *121*, 349–364. [CrossRef]
15. Yan, X.; Tang, G.; Wang, X. Bearing performance degradation assessment based on the continuous-scale mathematical morphological particle and feature fusion. *Measurement* **2022**, *188*, 110571. [CrossRef]
16. Muller, A.; Suhner, M.C.; Iung, B. Formalisation of a new prognosis model for supporting proactive maintenance implementation on industrial system. *Reliab. Eng. Syst. Saf.* **2008**, *93*, 234–253. [CrossRef]
17. Liang, S.; Zhang, Z.; Cui, L.; Zhong, Q. Dimensionality Reduction Method Based on Principal Component Analysis and Kernel Independent Component Analysis. *Syst. Eng. Electron. Technol.* **2011**, *33*, 2144–2148.
18. Li, Z.; Zhang, F.; Xiao, Y. Dynamic Blind Separation of Mechanical Fault Sources Based on CVA-ICA. *Chin. J. Mech. Eng.* **2015**, *51*, 24–29. [CrossRef]
19. Meng, Z.; Cai, L. Single-channel blind source separation based on EEMD sub-band extraction of relevant mechanical vibration signals. *Vib. Shock.* **2014**, *33*, 40–46+51.
20. Xu, Y.; Meng, Z.; Zhao, G. Research on Bearing Composite Fault Diagnosis Based on Double Tree Complex Wavelet Transform. *J. Instrum.* **2014**, *35*, 447–452.
21. Chen, G. Feature extraction and intelligent diagnosis of early faults of rolling bearings. *Chin. J. Aeronaut. Astronaut.* **2009**, 2.
22. Li, L.; Tang, M. Research on the diagnosis method of rolling bearing failure degree. *Bearing* **2009**, *4*, 42–46.
23. Chen, X.; Zhao, D.; Wang, G.; Xu, C. Fault monitoring of rolling bearing based on neural network. *Bearing* **2003**, *2*, 23–26.
24. Zhu, L.; Zhong, B.; Jia, M. Short-term analysis method of vibration signal and its application in mechanical fault diagnosis. *Chin. J. Vib. Eng.* **2000**, *13*, 400–405.
25. Wang, X. Research on Mechanical Fault Diagnosis Method Based on Kernel Function Method. Master's Thesis, Zhengzhou University, Zhengzhou, China, 15 February 2008.

sensors

MDPI

Article

Rolling Bearing Fault Diagnosis Based on WGWOA-VMD-SVM

Junbo Zhou [1], Maohua Xiao [1,*], Yue Niu [1] and Guojun Ji [2]

[1] College of Engineering, Nanjing Agricultural University, Nanjing 210032, China
[2] Essen Agricultural Machinery Changzhou Co., Ltd., Changzhou 213000, China
* Correspondence: xiaomaohua@njau.edu.cn

Abstract: A rolling bearing fault diagnosis method based on whale gray wolf optimization algorithm-variational mode decomposition-support vector machine (WGWOA-VMD-SVM) was proposed to solve the unclear fault characterization of rolling bearing vibration signal due to its nonlinear and nonstationary characteristics. A whale gray wolf optimization algorithm (WGWOA) was proposed by combining whale optimization algorithm (WOA) and gray wolf optimization (GWO), and the rolling bearing signal was decomposed by using variational mode decomposition (VMD). Each eigenvalue was extracted as eigenvector after VMD, and the training and test sets of the fault diagnosis model were divided accordingly. The support vector machine (SVM) was used as the fault diagnosis model and optimized by using WGWOA. The validity of this method was verified by two cases of Case Western Reserve University bearing data set and laboratory test. The test results show that in the bearing data set of Case Western Reserve University, compared with the existing VMD-SVM method, the fault diagnosis accuracy rate of the WGWOA-VMD-SVM method in five repeated tests reaches 100.00%, which preliminarily verifies the feasibility of this algorithm. In the laboratory test case, the diagnostic effect of the proposed fault diagnosis method is compared with backpropagation neural network, SVM, VMD-SVM, WOA-VMD-SVM, GWO-VMD-SVM, and WGWOA-VMD-SVM. Test results show that the accuracy rate of WGWOA-VMD-SVM fault diagnosis is the highest, the accuracy rate of a single test reaches 100.00%, and the accuracy rate of five repeated tests reaches 99.75%, which is the highest compared with the above six methods. WGWOA plays a good optimization role in optimizing VMD and SVM. The signal decomposed by VMD is optimized by using the WGWOA algorithm without mode overlap. WGWOA has the better convergence performance than WOA and GWO, which further verifies its superiority among the compared methods. The research results can provide an effective improvement method for the existing rolling bearing fault diagnosis technology.

Keywords: fault diagnosis; VMD; SVM; rolling bearing; WGWOA

Citation: Zhou, J.; Xiao, M.; Niu, Y.; Ji, G. Rolling Bearing Fault Diagnosis Based on WGWOA-VMD-SVM. *Sensors* **2022**, *22*, 6281. https://doi.org/10.3390/s22166281

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 21 July 2022
Accepted: 19 August 2022
Published: 21 August 2022

1. Introduction

Rolling bearing is the basic component of mechanical equipment and is used by most rotating machinery; it plays an important role in various fields of production [1,2]. Rolling bearing failure is likely to occur because it often operates under heavy load [3,4]. Statistics show that about 30% of mechanical faults in rotating machinery equipment using rolling bearings are related to bearing damage [5]. Fault diagnosis using vibration signals generated during its working process can reduce the probability of mechanical equipment accidents and provide reliable decision support for equipment later maintenance plans, which has high practical importance [6–8].

Rolling bearing vibration signal is nonlinear and complex due to many factors, so extracting effective information from it is particularly important [9–11]. Traditional methods mainly include time-domain and frequency-domain analyses. The former extracts and analyzes the statistical indexes of signals. The latter converts the signal into a frequency domain by Fourier transform and uses the fault frequency in the spectrum to make further analysis. However, the statistical characteristics of nonstationary complex signals in time and frequency domains must be analyzed and processed because they all change with

time [12]. Common time-frequency analysis methods mainly include wavelet transform, empirical mode decomposition (EMD), and local mean decomposition (LMD) [13–15]. EMD can adaptively decompose the signal into several eigenmode functions. Song et al. [16] proposed a novel bearing fault diagnosis based on EMD and improved Chebyshev distance and verified its accuracy and robustness by experiments. However, EMD is prone to problems, such as end-point effect and mode overlap [17,18]. Compared with EMD, LMD improves the number of iterations and the speed of operation, but it still cannot solve the problems of end-point effect. In 2014, Konstantin et al. [19] proposed a new variable scale processing method called variable mode decomposition (VMD) [20]. This method introduces a variational model and converts signal decomposition into an optimization problem of constrained model, which can avoid the end-point effect, restrain mode confusion, and has high decomposition efficiency. However, selecting the decomposition level and secondary penalty factor accurately is difficult in the application of VMD. Lin et al. [21] conducted gear fault diagnosis by using cuckoo search (CS) to optimize VMD and probabilistic neural network. The test results show that VMD can effectively avoid mode overlap, and the accuracy of this fault diagnosis method is 98.50%. However, the kurtosis index is selected as the fitness function of the CS algorithm, which results in unstable values of the decomposition layer.

With the development of machine learning and deep learning, the combination of intelligent learning algorithm and rolling bearing fault identification has become a hot research topic. The commonly used methods are artificial neural network (ANN), back-propagation neural network (BPNN), and support vector machine (SVM) [22–25]. As a classical algorithm in machine learning, SVM can solve the problems of ANN easily being fitted and BPNN easily falling into local optimum, so SVM is widely used in pattern recognition [26–28]. Van et al. [29] built a hybrid SVM model and applied it to bearing fault classification. They proved its superiority in terms of classification effect and training time by experiment. The selection of the structural parameters of SVM is difficult and directly affects its performance. Particle swarm optimization (PSO), whale optimization algorithm (WOA), gray wolf optimization (GWO), and other representative population algorithms are widely used in the optimization of SVM structural parameters [30,31] due to their advantages, such as good optimization performance and easy implementation. Garca et al. [32] proposed a PSO-SVM model to predict the remaining service life of aircraft engines. The test results show that the prediction accuracy is higher than that of the traditional PSO-SVM method. Dong et al. [33] presented a rolling bearing fault diagnosis model of GWO-SVM. The test results show that the GWO-SVM fault diagnosis model is more efficient than the SVM model. However, the PSO algorithm has the limitation of being trapped in local optimum [34], and the GWO algorithm has low optimization accuracy. Thus, these algorithms can still be improved. He et al. [35] developed an improved WOA algorithm called SWOA to optimize SVM and applied it to the prediction of soil moisture in maize. The test results show that the mean absolute error of the predicted results of this method is reduced from 0.87 to 0.67 compared with the optimized SVM of WOA, which proves the feasibility of the improved algorithm.

Although VMD improves in terms of the end-point effect, mode mixing, and other issues, selecting the decomposition layers and secondary penalty factors accurately is difficult. The SVM model is suitable for fault classification, but its performance largely depends on the constraints of core function parameters and penalty factors. At present, population optimization algorithm has certain parameter optimization capability, but its structure needs to be further improved to meet the actual needs.

We proposed a whale gray wolf optimization algorithm-VMD-SVM (WGWOA-VMD-SVM) for the fault diagnosis of rolling bearing. The vibration signal of rolling bearing is decomposed by VMD. A WGWOA algorithm based on WOA and GWO is presented. This algorithm is used to determine the best secondary penalty factor and decomposition layer number of VMD. The vibration signal of rolling bearing is decomposed into several components by using VMD optimized by WGWOA. The permutation entropies are

extracted as feature vectors. SVM is used as the rolling bearing fault diagnosis model, and the WGWOA algorithm is used to solve the optimal penalty factor and core function parameters. The optimized SVM model is trained in accordance with the extracted feature vector, and the test sample output is obtained. The fault diagnosis methods in this paper were comprehensively evaluated in terms of time-frequency signal, optimized fitness curve, and fault diagnosis accuracy to verify the feasibility and practicability of the proposed algorithm by two test cases.

2. Theoretical Basis

2.1. VMD

The essence of VMD is to decompose the vibration signal into several amplitude frequency-modulated signals by frequency domain iteration. For a group of complex vibration signals, the optimal variational model constructed by VMD can be decomposed into a series of intrinsic mode functions (IMFs) through multiple iterative calculations. In other words, the modal function $u_k(t)$, $k \in [1, B]$ is obtained with the minimum sum of B prediction bandwidths in the time series of the original signal.

Suppose a multifrequency signal F can be divided into k discrete time series $u_k(t)$ with limited bandwidth. Their corresponding central fundamental frequency band is $\omega_k(t)$. The spectrum obtained from $u_k(t)$ has sparse characteristics. The specific steps of bandwidth calculation are as follows.

The analytical signal and unilateral spectrum of the decomposed eigenmode function signal of each order are calculated by using Hilbert transform, which can be expressed as

$$\left[\delta(t) + \frac{j}{\pi \cdot t}\right] \cdot u_k(t) \tag{1}$$

Each modal signal is multiplied by an exponential term to make certain adjustments to its central frequency band:

$$\left\{\left[\delta(t) + \frac{j}{\pi \cdot t}\right] \cdot u_k(t)\right\} \cdot e^{-j \cdot w_k \cdot t} \tag{2}$$

The gradient norm of the demodulated signal is calculated, and the bandwidth of each modal signal is estimated, which can be expressed as

$$\left\| \partial_t \cdot \left\{\left[\delta(t) + \frac{j}{\pi \cdot t}\right] \cdot u_k(t)\right\} \cdot e^{-j \cdot w_k \cdot t} \right\|_2^2 \tag{3}$$

The center frequency and bandwidth obtained by the above equation are conditionally limited, that is, it should meet the requirements of minimizing the sum of the signal bandwidths of each IMF. Therefore, a constrained variational model should be developed.

$$\begin{cases} \left\| \partial_t \cdot \left\{\left[\delta(t) + \frac{j}{\pi \cdot t}\right] \cdot u_k(t)\right\} \cdot e^{-j \cdot w_k \cdot t} \right\|_2^2 \\ s.t \sum_{k=1}^{K} u_k = f \end{cases} \tag{4}$$

where ω_k is the frequency center of each IMF; u_k is the kth IMF; f is the original signal.

The quadratic penalty factor method and the Lagrange function multiplier method are introduced to transform the above equation into an unconstrained variational problem and to obtain its optimal solution. The augmented Lagrange function is

$$L(u_k, \omega_k, \lambda) = \alpha \cdot \sum_{k=1}^{K} \left\| \partial_t \cdot \left\{\left[\partial(t) + \frac{j}{\pi \cdot t}\right] \cdot u_k(t)\right\} \cdot e^{-j \cdot \omega_k \cdot t} \right\|_2^2 + \left\| f(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_{k=1}^{K} u_k(t) \right\rangle \tag{5}$$

The alternating direction method of multipliers is introduced to search the saddle point of the variational problem. The center frequency and bandwidth of each IMF signal can be updated:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \cdot |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega} \tag{6}$$

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum\limits_{i \neq k} \hat{u}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2 \cdot \alpha \cdot (\omega - \omega_k)^2} \tag{7}$$

where $\hat{u}_k^{n+1}(\omega)$ is the filtering result of residual quantity $\hat{f}(\omega) - \sum\limits_{i \neq k} \hat{u}_i(\omega)$; ω_k^{n+1} is the power spectrum center of gravity of the current mode; and the real part can be obtained by inverse fast Fourier transformation of $\hat{u}_k(t)$.

The decomposition steps of VMD are as follows:

Step 1: Initialize parameter u_k, ω_k, λ, α, and N;

Step 2: $N = N + 1$, and the VMD algorithm is used for iterative calculation;

Step 3: The value of k is continuously superimposed from 1 to k, u_k and ω_k are continuously updated by using Equations (5) and (6), respectively, and k is the total amount of IMF finally decomposed;

Step 4: Update λ in accordance with the following equation:

$$\lambda^{n+1} = \lambda^n + \tau \cdot (f - \sum_k \hat{u}_k^{n+1}) \tag{8}$$

Step 5: Give the judgment accuracy $\varepsilon > 0$, and repeat steps (3) and (4) until the termination conditions of the following equation are met:

$$\sum_k \frac{\|u_k^{n+1} - u_k^n\|_2^2}{\|u_k^n\|_2^2} < \varepsilon \tag{9}$$

VMD can effectively avoid the phenomenon of modal aliasing and can perform effective signal analysis to extract differentiated eigenvalues due to its strong robustness. Therefore, the signal after VMD can effectively describe the characteristics of fault signals.

2.2. WGWOA

GWO is an algorithm proposed by Mirjalili et al. [36]. Its basic principle is to imitate the population system of gray wolves and divide them into α, β, δ, and γ. Gray wolf γ accepts gray wolves during the hunting of α, β, and δ. The process of the gray wolf algorithm can be divided into three stages: encirclement, pursuit, and attack [37,38], and the specific steps are as follows:

Step 1: Surround prey

In the GWO algorithm, each gray wolf individual realizes prey encirclement in accordance with the following equation:

$$D = \left| C \cdot x_{p(t)} - x_t \right| \tag{10}$$

$$x_{t+1} = x_{p(t)} - A \cdot D \tag{11}$$

where D is the Euclidean distance between the wolf individual and its prey; $X_{(p(t))}$ is the location of the prey; X_t is the individual position of wolf before the start of the enclosure process; X_{t+1} is the individual position of the gray wolf at the end of the enclosure process.

The calculation equations of variable coefficients A and C are as follows:

$$A = 2 \cdot a \cdot r_1 - a \tag{12}$$

$$C = 2 \cdot r_2 \tag{13}$$

where a is the contraction factor, which decreases linearly from 2 to 0; r_1 and r_2 are two different [0, 1] random numbers.

Step 2: Hunt prey

After surrounding the prey, gray wolves α, β, and δ are three potential solutions. All the individuals in the wolf pack are in the GWO algorithm. α, β, and δ are led by prey hunting, and each gray wolf individual follows the following equations for pursuing:

$$D_q = \left| C_l \cdot x_j - x_{f(t)} \right| \tag{14}$$

$$x_l = x_q - A_l \cdot D_q \tag{15}$$

$$x_{f(t+1)} = \frac{\sum x_l}{3} \tag{16}$$

where q takes α, β, and δ; l taken 1, 2, and 3; Dq is the Euclidean distance between the q wolf and the gray wolf. x_l is the distance from the individual gray wolf to the q wolf; $x_{f(t)}$ is the individual position of the gray wolf before the start of the chase; $x_{f(t+1)}$ is the individual position of the gray wolf after the end of catching; A_l and C_l of the coefficient of variation are determined by using Equations (12) and (13), respectively.

Step 3: Attack prey

When the prey is surrounded by a pack of wolves, the pack begins attacking the prey. When a decreases linearly from 2 to 0, the range of A is $[-a, a]$, as shown in Equation (12). When $|A| < 1$, the gray wolf is attacking its prey. When $|A| > 1$, the gray wolf leaves the wolf pack to find the next prey and expand the entire wolf pack search capability.

The GWO algorithm is nongreedy in nature, so it has good global optimization ability and is not easy to fall into local optimum. However, the GWO algorithm only uses straight-line hunting to catch prey, which restricts its search range and accuracy, resulting in slow convergence speed and poor local optimization capability. Therefore, the manner of the gray wolf algorithm to chase prey needs to be improved.

WOA is an algorithm that simulates whale predation in nature. It is divided into three stages: surround prey, bubble attack, and search-and-prey [39]. During the bubble attack phase, each individual chases its prey in a shrink enclosure with a 50% probability, similar to the way in which the individual chases its prey in the GWO algorithm (Equation (15)) and spirals its prey with a 50% probability. The whale algorithm uses the following methods of bubble attack:

$$D_w = |x_{wbest} - x_w| \tag{17}$$

$$x_{w+1} = \begin{cases} x_{wbest} - A_w \cdot D_w & r_3 < 0.5 \\ x_{wbest} + D_w e^{b \cdot R} \cdot \cos(2 \cdot \pi \cdot R) & r_3 \geq 0.5 \end{cases} \tag{18}$$

where D_w is the Euclidean distance between the individual whale and the best individual whale; x_{wbest} is the position of the best individual whale; x_w is the individual position of the whale before bubble attack; x_{w+1} is the individual position of the whale after bubble attack; b is the logarithmic spiral shape constant; R is a random number between $[-1, 1]$; r_3 is a random number between [0, 1]; the variable coefficient A_w is determined in the same manner as Equation (12).

Inspired by the bubbling attack mode of the WOA algorithm, the WGWOA algorithm is proposed. The stages of enclosing and attacking prey in this algorithm are consistent with the GWO algorithm, and the manner of chasing prey is as follows:

$$D_q = \left| C_l \cdot x_j - x_t \right| \tag{19}$$

$$x_l = \begin{cases} x_q - A_l \cdot D_q & r_3 < 0.5 \\ x_q + D_q e^{b \cdot R} \cdot \cos(2 \cdot \pi \cdot R) & r_3 \geq 0.5 \end{cases} \tag{20}$$

$$x_{t+1} = \frac{\sum x_l}{3} \tag{21}$$

where x_t is the position before the start of the individual chase in the WGWOA algorithm; x_{t+1} is the position before the start of the individual chase.

Equations (19)–(21) show that the WGWOA algorithm still follows the wolf-led strategy of the gray wolf algorithm, retains the nongreedy algorithm with strong global optimization ability, and introduces the bubble attack mode of the WOA algorithm, which improves the population diversity, local optimization ability, and convergence performance. Therefore, the WGWOA algorithm considers the global and local optimization performance of the algorithm.

2.3. VMD Optimized Based on the WGWOA Algorithm

In the VMD process, the quadratic penalty factor σ and the number of IMF components K have a great influence on its decomposition results. The values of σ and K depend on the empirical parameters in the literature, which to a large extent has a tentative problem, and their applicability is limited. If the two parameters are not selected well, the signal will not be well-decomposed, resulting in over decomposition or under decomposition, which affects the extraction and judgment of important information.

Therefore, the VMD algorithm should be improved so that the appropriate σ and K values can be selected to realize the correct decomposition of the vibration signal of the rolling bearing. In this paper, the WGWOA algorithm is used to optimize the parameters σ and K of the VMD algorithm, and adaptive selection is performed to determine the best combination of parameters [σ, K].

Permutation entropy is a dimensionless index used to characterize the complexity of signal sequence and has many advantages, such as low requirement for sequence length and strong robustness. Therefore, it is widely used in condition monitoring, fault diagnosis, and signal detection of mechanical equipment [40]. Therefore, the permutation entropy of each component of VMD is used as the fitness function in the optimization of WGWOA algorithm due to the characteristics of permutation entropy.

Assuming a signal of length L: $\{y(i), i = 1, 2, \ldots, L\}$, the permutation entropies are calculated as follows:

Step 1: Spatial reconstruction

$$\begin{bmatrix} x(1) & x(1+\tau) & x[1+(m-1)\cdot\tau] \\ x(2) & x(2+\tau) & x[2+(m-1)\cdot\tau] \\ x(z) & x(x+\tau) & x[z+(m-1)\cdot\tau] \\ & \cdots & \\ x(\kappa) & x(\kappa+\tau) & x[\kappa+(m-1)\cdot\tau] \end{bmatrix}, \; z = 1, 2, \ldots, \kappa[\kappa = L-(m-1)\cdot\tau] \tag{22}$$

where m is the embedding dimension; τ is the delay time.

Step 2: Reconstruct the zth reconstructed component $x(z), x(z+\tau), \ldots, x[z+(m-1)\cdot\tau]$ in ascending order. The values of $z_1, z_2, \ldots, z_m$ indicate the index of the column in which each element in the reconstructed component is located. A set of symbolic sequences can be obtained for each row of the reconstruction matrix of any time series $y(i)$ reconstructed from the phase space.

$$S(\xi) = (z_1, z_2, \ldots, z_m), \quad \xi = 1, 2, \ldots, \theta(\theta \le m!) \tag{23}$$

Step 3: m-dimensional phase space is mapped to $m!$, and $S(\xi)$ is only one of the different sequences of symbols. If the occurrence probability of each sequence of symbols is recorded as $P_1, P_2, \ldots, P_\theta$, then the permutation entropy is calculated as follows:

$$PE = -\sum_{i=1}^{\theta} P_z \cdot \ln P_z \tag{24}$$

The process of VMD optimization based on the WGWOA algorithm is as Figure 1:

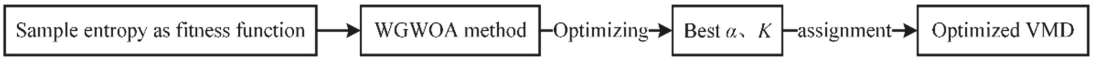


Figure 1. Flowchart of VMD optimization based on the WGWOA algorithm.

2.4. Fault Diagnosis Model Based on Optimized SVM

SVM is a machine learning method based on statistical learning theory. Its algorithm is characterized by maximizing the interval and it can find the optimal classification hyperplane [41] that separates different types of sample data and has the maximum classification interval. SVM can map input sample space to high-dimensional feature space by means of "core mapping," overcome the problems of "dimension disaster" and "overfitting" in traditional machine learning model, and show great advantages in solving small sample, nonlinearity, and high-dimensional identification [42]. Therefore, SVM is used to construct a fault pattern recognition model. SVM needs to train the test data to realize fault identification, so the characteristic value of bearing vibration signal should be extracted to construct the data set. Based on the advantages of permutation entropy described above, the permutation entropy of each component after WGWOA-VMD decomposition is extracted to form a feature vector.

The penalty factor c and parameter g of radial basis core function have great influence on the performance of SVM during training. Therefore, the two parameters c and g of SVM are optimized by using the proposed WGWOA algorithm. A fault diagnosis model based on the optimized SVM was proposed. The flow chart of the algorithm is shown in Figure 2.

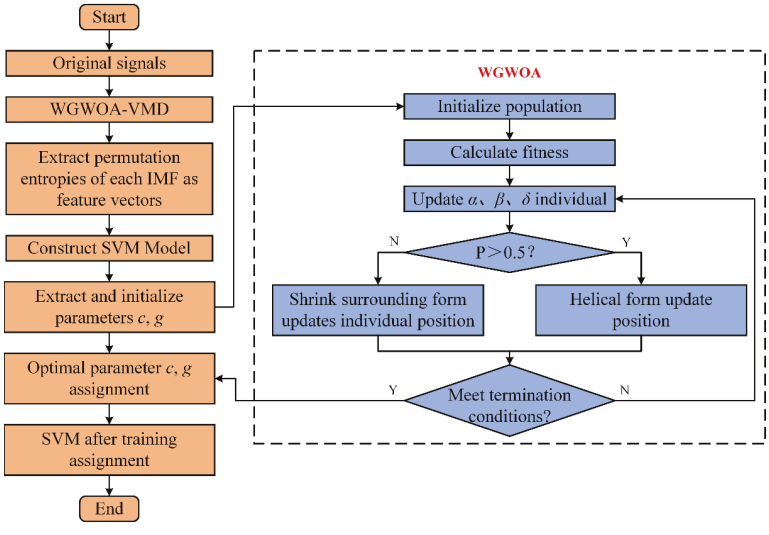


Figure 2. Flowchart of fault diagnosis based on the optimized SVM.

3. Fault Diagnosis of Rolling Bearing Based on WGWOA-VMD-SVM

The rolling bearing fault signal is processed and recognized through signal processing, feature extraction, and pattern recognition. The general research route and basic theory are shown in Figure 3. In signal processing, the vibration signal is decomposed by VMD, the WGWOA algorithm is proposed to calculate the parameters in VMD, and σ is optimized with K. In feature extraction, the permutation entropy of each IMF decomposed by WGWOA-VMD is extracted to form the characteristic vector of vibration signal. In the aspect of fault pattern recognition, the characteristic vectors of each signal are inputted into

the SVM model for fault diagnosis and classification, and the WGWOA algorithm is used to optimize the important parameters c and g of SVM.

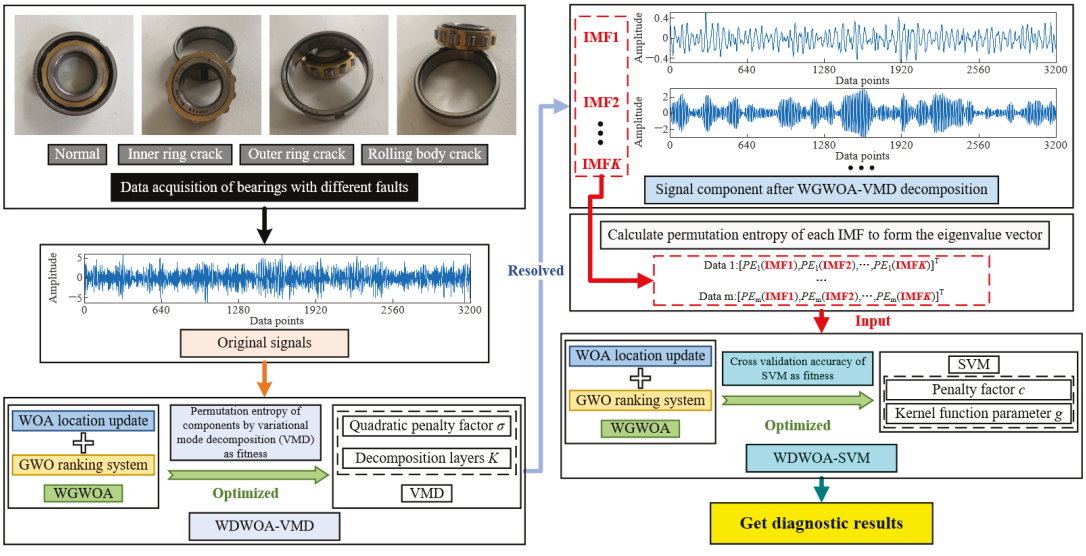


Figure 3. Fault diagnosis of rolling bearing based on WGWOA-VMD-SVM.

The specific steps of rolling bearing fault diagnosis based on WGWOA-VMD-SVM are as follows:

Step 1: The fault states of normal rolling bearing, inner ring crack, outer ring crack, and rolling element crack are sampled many times;

Step 2: Taking the permutation entropy of each component VMD decomposed of signal samples as the fitness function, the WGWOA algorithm is used to decompose the input parameters of VMD in each fault condition by the number of levels K and the quadratic penalty factor σ. At the same time, input the vibration signal training samples, perform VMD, and obtain K eigenmode function components;

Step 3: The permutation entropy of K modes is extracted as the sample eigenvector, and the eigenvalue matrix is constructed;

Step 4: Taking the accuracy of SVM cross-validation as the fitness function, the proposed WGWOA algorithm is used to optimize the SVM parameters c and g;

Step 5: Input the test samples into the trained SVM, obtain the diagnostic results, and verify the training effect.

4. Experimental Research Based on Public Data Set

4.1. Test Data Acquisition

The bearing vibration signals collected by Case Western Reserve University were used as the experimental data, which was based on the test bench shown in Figure 4. In this experiment, the bearings at the motor drive end and the fan end were taken as the diagnostic objects, and the single point damage was introduced on the inner ring, outer ring, and roller of the test bearing by EDM to simulate three kinds of bearing faults. The damage size was 0.1778 mm, 0.3556 mm, and 0.5334 mm, respectively, and then the signals were collected by the acceleration sensor under different working conditions.

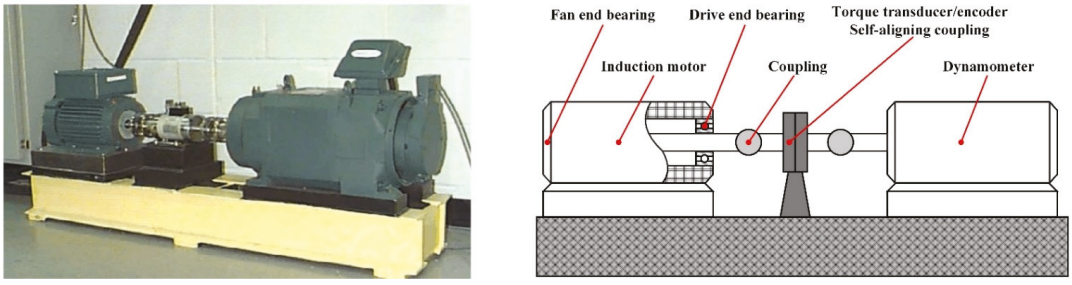


Figure 4. Rolling bearing fault simulation experimental device.

In this test, the vibration signals of the rolling bearing at the driving end under normal condition, inner ring fault, outer ring fault, and roller fault with diameters of 0.1778 mm, 0.3556 mm, and 0.5334 mm, respectively, and loads of 0HP (speed 1796 r·min^{-1}), 1HP (speed 1772 r·min^{-1}), and 2HP (speed 1750 r·min^{-1}) were analyzed. The sampling frequency was 12 kHz, the time domain diagram of some vibration signals of the test are shown in Figure 5.

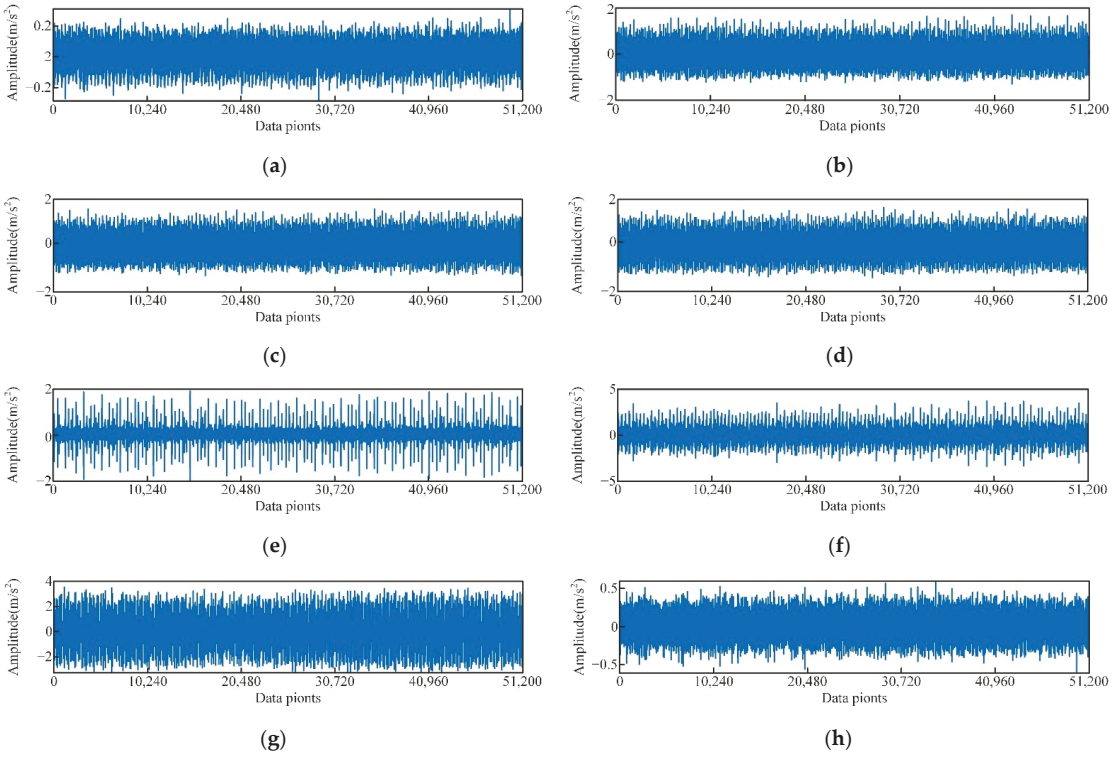
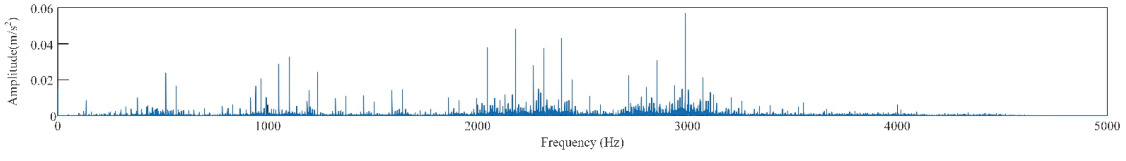
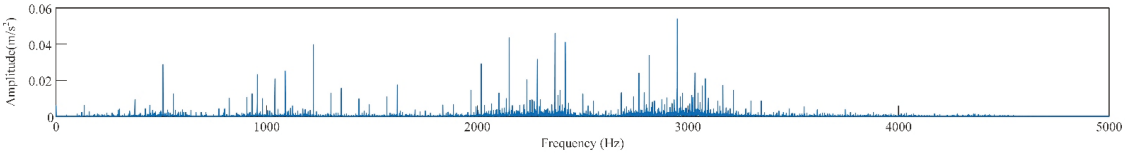


Figure 5. Time domain diagram of vibration signals of different types of rolling bearings, (**a**) 0HP load normal, (**b**) 0HP load inner ring fault diameter is 0.1778 mm, (**c**) 1HP load inner ring fault diameter is 0.1778 mm, (**d**) 2HP load inner ring fault diameter is 0.1778 mm, (**e**) 0HP load inner ring fault diameter is 0.3556 mm, (**f**) 0HP load inner ring fault diameter is 0.5334 mm, (**g**) 0HP load outer ring fault diameter is 0.1778 mm, (**h**) 0HP load rolling element fault diameter is 0.1778 mm.

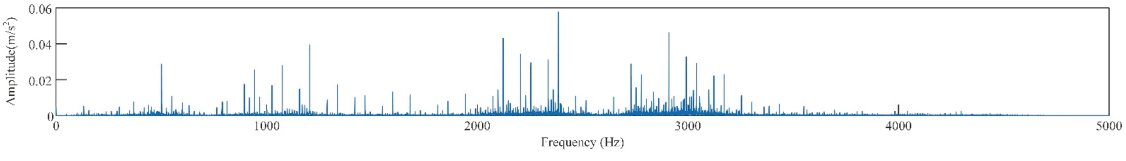
As shown in Figure 5b–d, for the same fault type, the signal discrimination of rolling bearing under different loads is small. Comparing the spectrograms of corresponding signals shown in Figure 5b–d (Figure 6), the spectrograms of the three signals are also quite similar. This is because the rotational speeds of the experimental data under different loads are similar and the frequency of the characteristic pulse occurrence is very small, so the spectrum peaks of the fault characteristic frequency in the spectrogram are not significantly different either. As shown in Figure 5b,e, and f, there are significant differences in time domain waveforms of the vibration signal of the rolling bearing with different fault diameters [43]. Compared with normal bearings, the amplitude of faulty bearings and obvious periodic vibration impact are obvious, as shown in Figure 5a,b,g,h. In the spectrum diagram (Figure 7), the spectrum of normal bearing vibration signal is relatively single from Figure 7a, and the energy mainly concentrates in the low frequency band. Figure 7b,c shows that the energy of inner and outer ring fault vibration signal mainly concentrates in the middle frequency band, and the low frequency is reflected in the spectrum. It can be seen from the failure of the rolling body in Figure 7d. As shown in Figure 7d, when a rolling element fails, it is accompanied by more prominent energy in both low and medium frequency bands, The signal is also rather cluttered.



(**a**)



(**b**)



(**c**)

Figure 6. Frequency domain diagram of vibration signal of rolling bearing with inner ring fault under different loads, (**a**) 0HP, (**b**) 1HP, (**c**) 2HP.

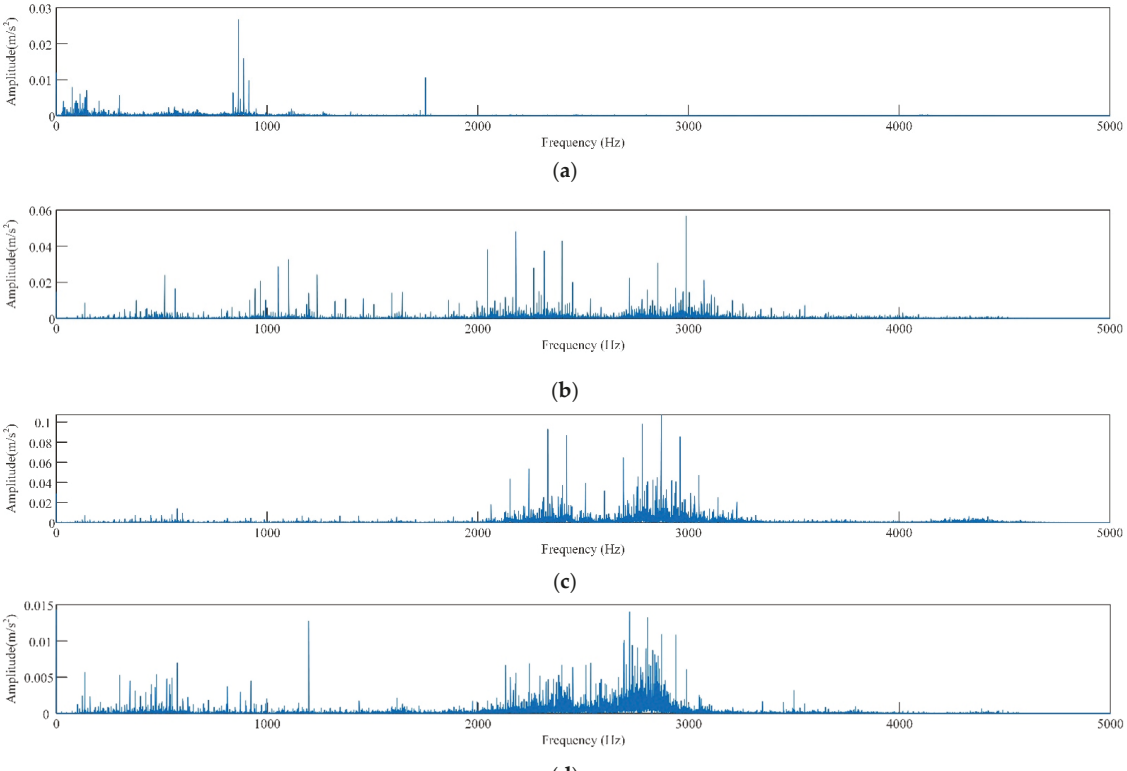


Figure 7. Frequency domain diagram of rolling bearing vibration signals of different fault types, (**a**) normal, (**b**) inner ring damaged, (**c**) outer ring damaged, (**d**) rolling body damaged.

Although the vibration signals of different faults are different, these signals are only individual ideal signals. In fact, the waveforms of some states are very similar and difficult to distinguish. Therefore, it is necessary to further separate and extract the characteristics of vibration signals by mode decomposition of each signal.

4.2. Signal Processing and Feature Extraction

The test samples are set as follows: the vibration signals of 51,200 data points of each type are collected. Because the load of the rolling bearing will change under actual conditions, the vibration signals of three loads under the same fault type were randomly combined according to the load type to detect whether the method in this paper can identify the same fault under different loads. According to this, a total of 153,600 data points were obtained for each new combination signal. Vibration signals of 2048 data points were VMD decomposed, and the permutation entropy of each component was extracted as the characteristic vector. A total of 75 samples were obtained for each fault. Then, 45 samples were randomly selected as the training sample set, and the remaining 30 samples were used as the test sample set. Each test sample set is as shown in Table 1.

Table 1. Sample Settings.

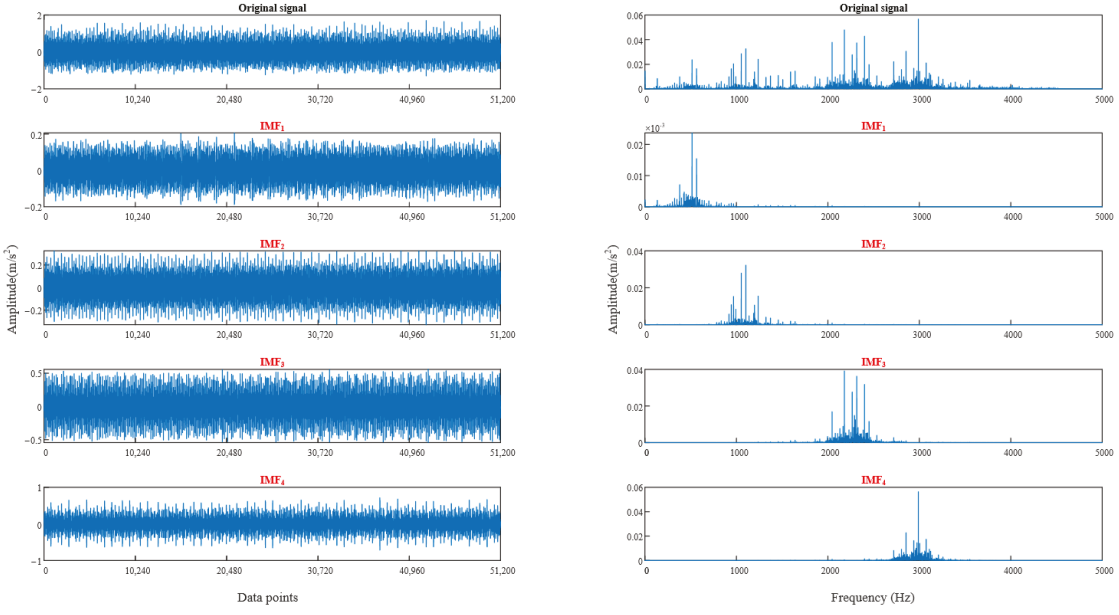
Fault Types	Load/(hp)	Number of Training Samples	Number of Test Samples	Sample Number
Normal	0	45	15	1
	1			
	2			
Inner ring fault (fault diameter 0.1778 mm)	0	45	15	2
	1			
	2			
Inner ring fault (fault diameter 0.3556 mm)	0	45	15	3
	1			
	2			
Inner ring fault (fault diameter 0.5334 mm)	0	45	15	4
	1			
	2			
Outer ring fault (fault diameter 0.1778 mm)	0	45	15	5
	1			
	2			
Outer ring fault (fault diameter 0.3556 mm)	0 1 2	45	15	6
Outer ring fault (fault diameter 0.5334 mm)	0	45	15	7
	1			
	2			
Rolling element fault (fault diameter 0.1778 mm)	0	45	15	8
	1			
	2			
Rolling element fault (fault diameter 0.3556 mm)	0	45	15	9
	1			
	2			
Rolling element fault (fault diameter 0.5334 mm)	0	45	15	10
	1			
	2			

Taking each component permutation entropy of vibration signal VMD decomposed as the fitness function, the parameters σ and K in VMD were determined by WGWOA. The optimal parameters determined by the algorithm are shown in Table 2, the parameters of different vibration signals obtained by the algorithm are relatively centralized, the decomposition layers K are all four, and the secondary penalty factor σ fluctuates slightly around 2000. In order to guarantee the optimization effect and the universality of the method, when the fault type of vibration signal to be diagnosed is unknown, reasonable signal decomposition is carried out. In this paper, the best combination of parameters was determined by obtaining the average value of VMD optimal parameters of σ and K different vibration signals [1996.20, 4].

Table 2. Optimal parameter solutions.

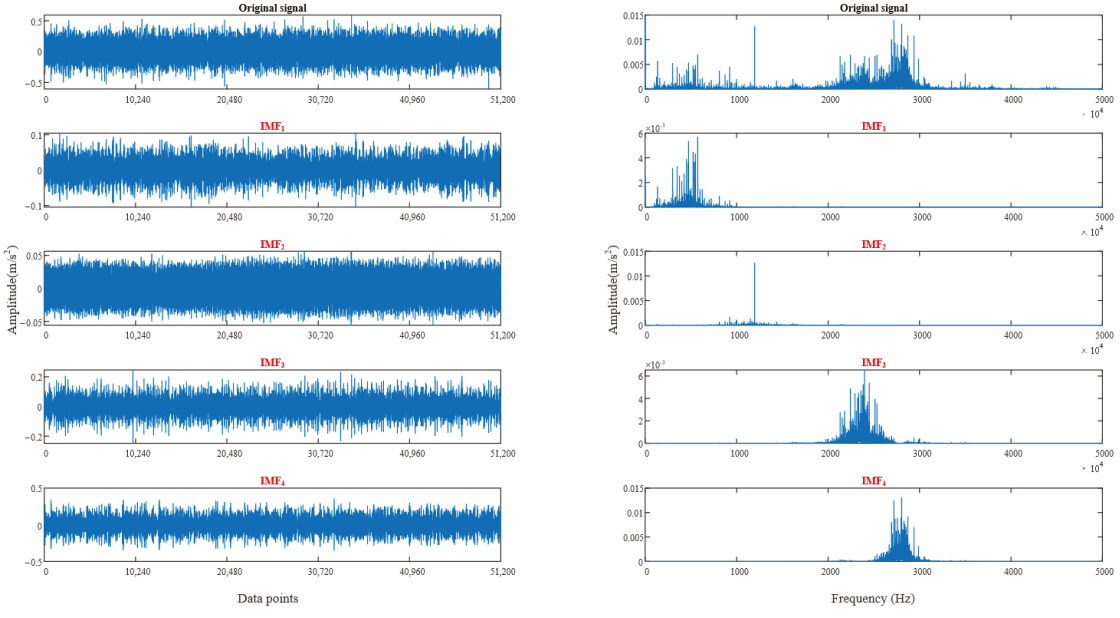
Fault Types	Optimum Solutions	
	σ	K
Normal	2012	4
Inner ring fault (fault diameter 0.1778 mm)	1999	4
Inner ring fault (fault diameter 0.3556 mm)	1982	4
Inner ring fault (fault diameter 0.5334 mm)	2003	4
Outer ring fault (fault diameter 0.1778 mm)	1996	4
Outer ring fault (fault diameter 0.3556 mm)	1988	4
Outer ring fault (fault diameter 0.5334 mm)	1999	4
Rolling element fault (fault diameter 0.1778 mm)	2007	4
Rolling element fault (fault diameter 0.3556 mm)	1987	4
Rolling element fault (fault diameter 0.5334 mm)	1989	4
Optimum parameter combination	1996.20	4

In order to verify the rationality of selecting the best parameter combination, Figure 8 shows the time-domain waveform and spectrum of bearing vibration signal after using the best parameter VMD (only two signals are listed here due to the length of the article). From the spectrum diagram, the vibration signal of each component can accurately reflect the characteristics of the original signal after using the best parameter combination VMD, and there is no modal aliasing, which proves the feasibility of WGWOA-VMD.



(**a**)

Figure 8. *Cont.*



(**b**)

Figure 8. Time domain waveform and spectrum diagram of vibration signal decomposed by WGWOA-VMD, (**a**) inner ring fault, (**b**) rolling body fault.

4.3. Fault Diagnosis Results and Comparative Analysis

SVM is used as the fault diagnosis model, and the correct cross-validation is the fitness during SVM training. The WGWOA algorithm is used to optimize the parameters *c* and *g* of its SVM, and the final optimal *c*, *g* solution combination is [16.58, 3.83]. Figure 9 is the fitness curve of the SVM training process. The SVM is trained with training samples, and the test samples are input into the trained SVM to output the diagnosis results.

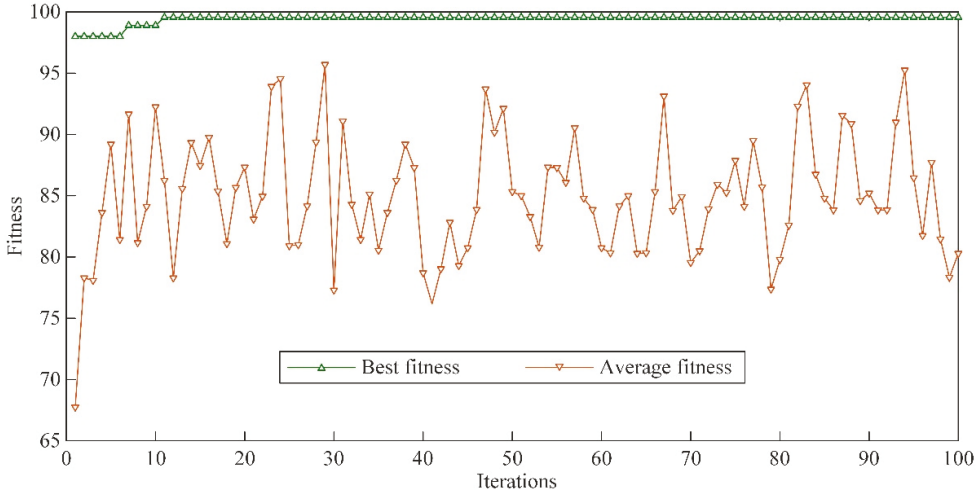
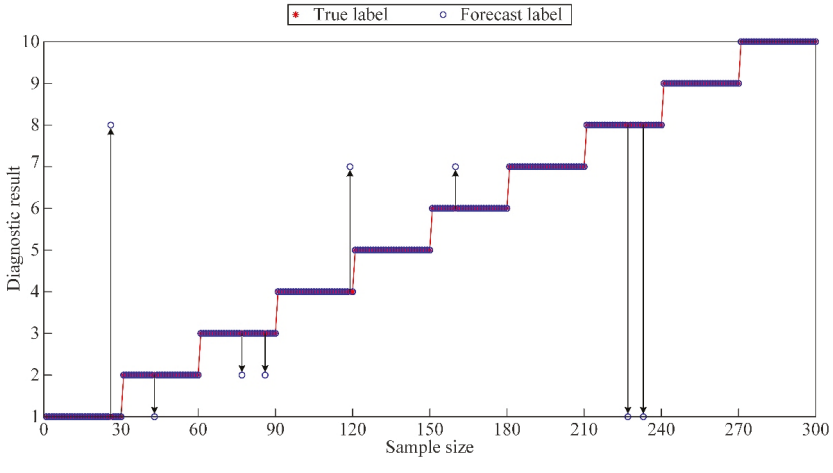
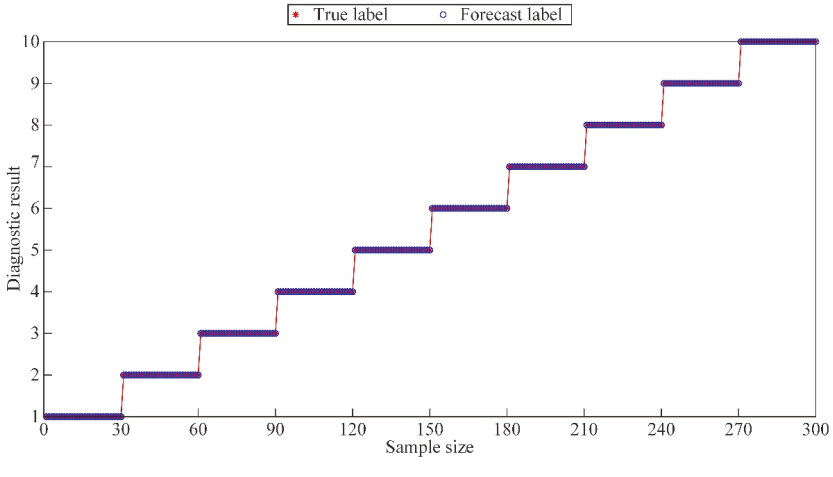


Figure 9. Fitness curve of SVM training process.

In order to preliminarily verify the feasibility of the fault diagnosis method in this paper, the existing fault diagnosis method combining VMD and SVM was used for a comparative test. As shown in Figure 10, the fault diagnosis accuracy rate of WGWOA-VMD-SVM method in this paper reaches 100.00%, and the accuracy rate of VMD-SVM fault diagnosis method is 97.33%. This is because WGWOA-SVM adopts the WGWOA algorithm to optimize the parameters of VMD and enhance the effect of signal decomposition. At the same time, the WGWOA algorithm is used to optimize the parameters of SVM and improve the recognition ability of the SVM model. In order to avoid contingency, five repeated tests were carried out for the two fault diagnosis methods. The experimental results are shown in Table 3. The five fault diagnosis rates of WGWOA-VMD-SVM method are all 100.00%, indicating that the fault diagnosis method in this paper has strong stability.



(**a**)



(**b**)

Figure 10. Fault diagnosis results of different methods, (**a**) VMD-SVM, (**b**) WGWOA-VMD-SVM.

Table 3. Diagnostic accuracy of different methods.

Methods	Accuracy (%)					
	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Average
VMD-SVM	97.33	96.00	98.66	94.00	97.33	96.66
WGWOA-VMD-SVM	100.00	100.00	100.00	100.00	100.00	100.00

5. Laboratory Test Research

5.1. Sources of Test Data

The bearing life cycle test platform of a mechanical transmission system independently developed by Nanjing Agricultural University was used for the test. As shown in Figure 11a,b, it is mainly composed of an integrated console, bearing pedestal, servo electric cylinder, motor, data acquisition card, acceleration sensor PCB35A26, temperature sensor, and pressure sensor. The motor speed and load are adjusted by the integrated console. During the test vibration signal collection, the motor drives the shaft to rotate, and the fault bearing is installed in the bearing seat of the shaft. The data acquisition card and the acceleration sensor are used to collect the bearing vibration data. The magnet at the bottom of the acceleration sensor is adsorbed in the radial direction of the bearing seat to be tested. After the test bench runs for 2 min, the running state is stable. The computer end acquisition software is used to start collecting the bearing vibration signal. App 2kN load to the motor through the load knob on the console and set the speed and sampling frequency to 1500 r·min^{-1} and 16 kHz, respectively. The bearing used in the test is a cylindrical roller bearing with the model of N205EM. The specific parameters are shown in Table 4. Regular cracks with a width of 0.2 mm and a depth of 0.5 mm were machined by EDM to simulate the fault bearing. The test bearing types include normal bearing, inner ring cracked bearing, outer ring cracked bearing, and rolling element cracked bearing (as shown in Figure 11c–f, and the quantity is one for each bearing).

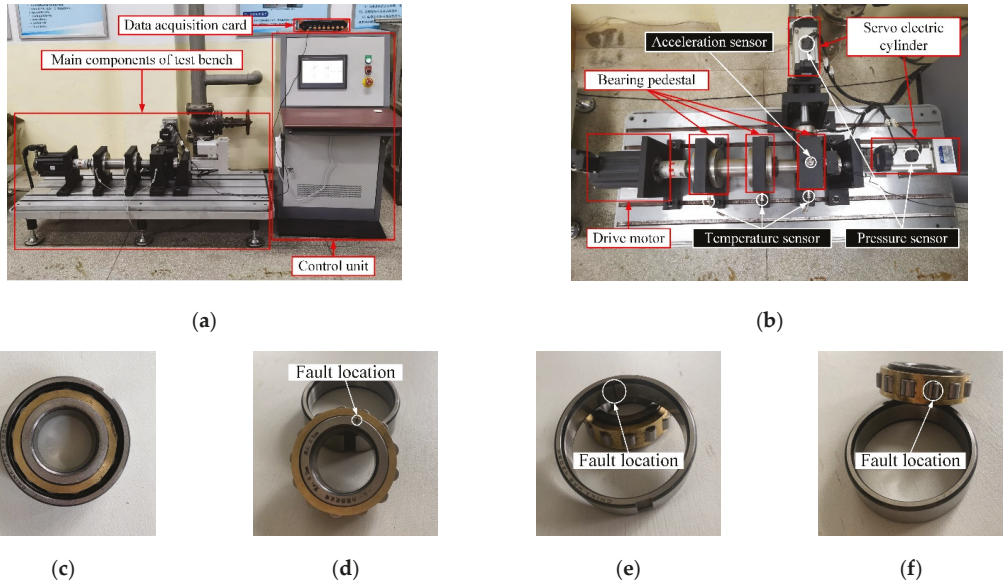


Figure 11. Test materials, (**a**) general layout of test stand, (**b**) schematic of the main structure of test stand, (**c**) normal bearings, (**d**) inner ring cracked bearings, (**e**) outer ring cracked bearings, (**f**) roller cracked bearings.

Table 4. Specifications and parameters of test bearings.

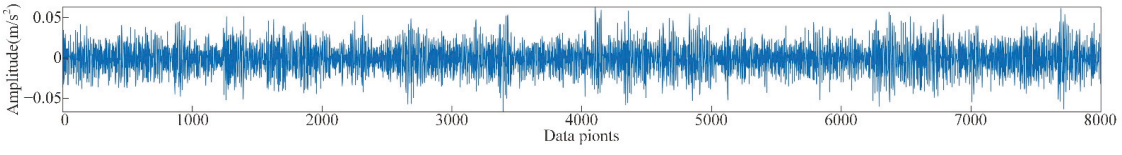
Types	Specifications	Outer Diameter/mm	Inside Diameter/mm	Thickness/mm	Rollers Number	Roller Diameter/mm	Pitch/mm	Contact Angle/°
Cylindrical roller bearing	N205EM	52	25	15	13	6.5	38.5	0

5.2. Preprocessing of Test Data and Feature Extraction

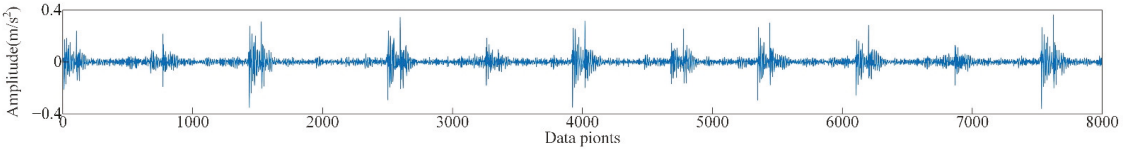
5.2.1. Data Preprocessing

The test data in this paper were set as follows: vibration signals of 80,000 data points in 5 s of each fault were collected, vibration signals of 1600 data points in 0.1 s were decomposed by VMD, permutation entropy of each IMF after VMD were extracted to form characteristic vectors, and 50 sets of data were obtained for each fault. Given that four types of faults are found in this test, 200 sets of data were set up to randomly divide the sample data sets of each fault condition in accordance with the proportion, avoiding phenomena, such as model fitting. Thirty groups (120 groups) of the bearing data of each state were used as training data for SVM, and the remaining 20 groups (80 groups) were used as test data for SVM.

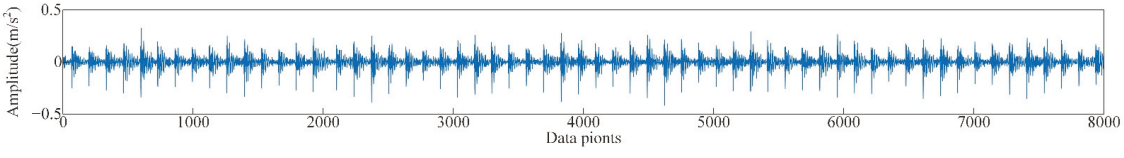
As shown in Figure 12, the vibration signal within 0.5 s (8000 data points) was collected for this experiment. As shown in the diagram, the vibration signal of normal bearing (Figure 12a) is relatively stable, with small amplitude and no large pulse. The vibration signals of faulty bearings (Figure 12b–d) differ from those of normal bearings. The time domain waveforms of fault bearing vibration signals have a larger amplitude and larger periodic vibration impact, various fault time domain diagrams are different, but it is not easy to determine the specific fault characteristics. Therefore, the characteristics of each vibration signal should be further separated and extracted by signal mode decomposition.



(**a**)

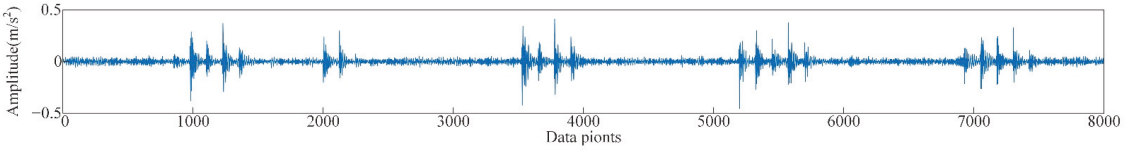


(**b**)



(**c**)

Figure 12. *Cont.*

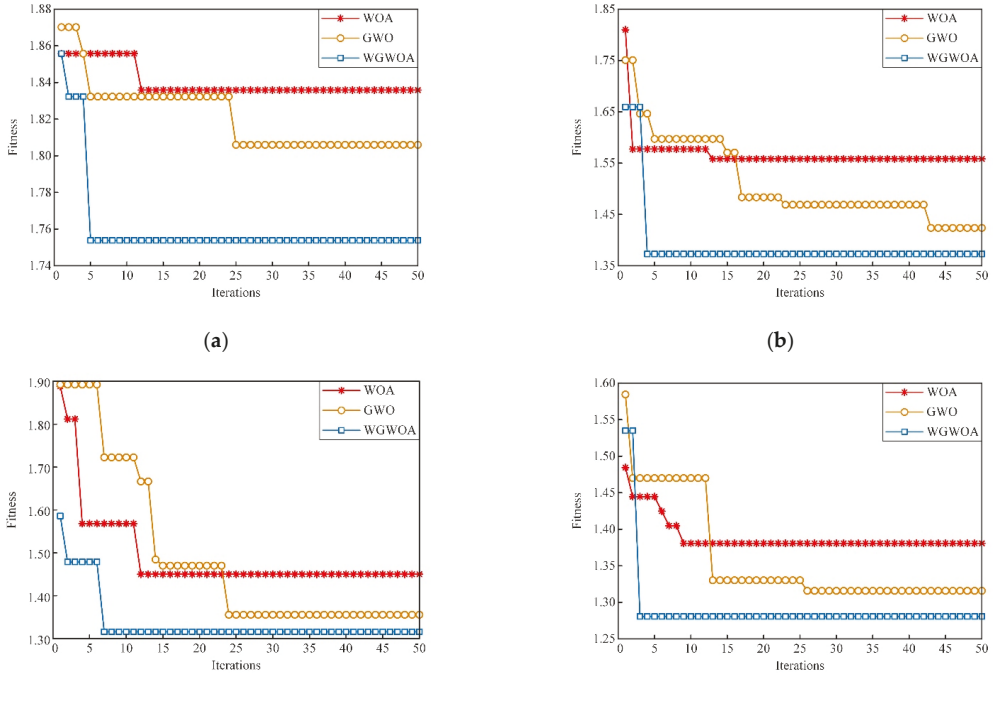


(**d**)

Figure 12. Time-domain waveform of the vibration signals of bearings with different faults, (**a**) normal bearings, (**b**) inner ring cracked bearings, (**c**) outer ring cracked bearings, (**d**) roller cracked bearings.

5.2.2. Signal Decomposition and Feature Extraction Based on WGWOA-VMD

Take the first component of different types of signals decomposed by VMD as an example. The fitness curves of the WGWOA, WOA, and GWO algorithms in VMD optimization are compared, and the number of iterations of the algorithm is 50 to verify the feasibility of WGWOA algorithm in optimizing the VMD parameters. As shown in Figure 13, three different algorithms are used to optimize the fitness curves of VMD.



(**a**) (**b**)

(**c**) (**d**)

Figure 13. Different algorithms to optimize the fitness curve of VMD, (**a**) normal bearings, (**b**) inner ring cracked bearings, (**c**) outer ring cracked bearings, (**d**) roller cracked bearings.

From Figure 13, the WOA algorithm has the highest adaptability in the bearing signal decomposition of four fault types, and the solution may be local optimum, which proves that the WOA algorithm is easy to fall into local optimum. The adaptability of GWO algorithm is lower than that of the WOA algorithm, which indicates that GWO has a stronger global optimization ability than WOA, but it converges only when the number

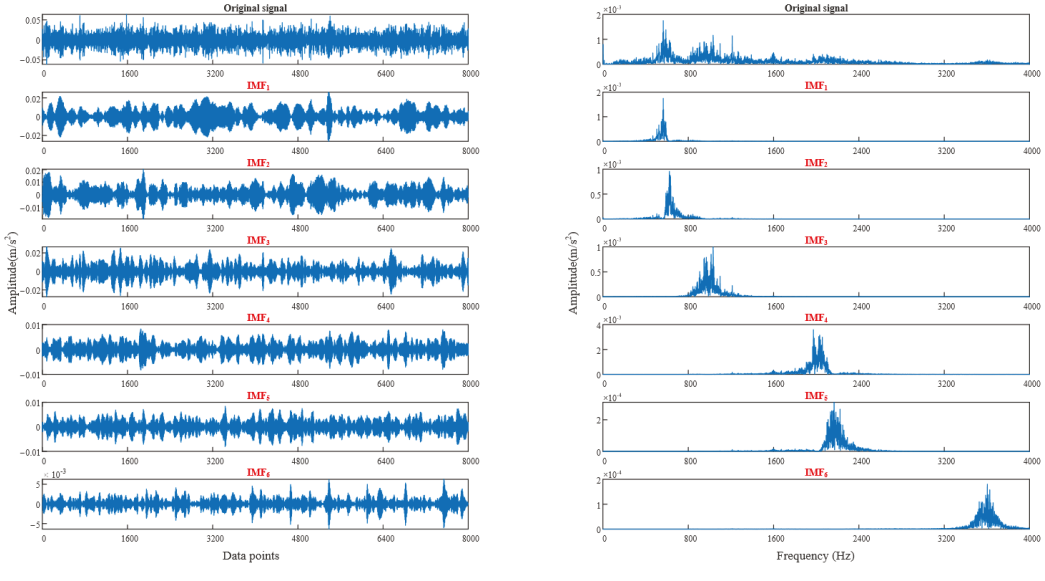
of iterations is higher, and its convergence ability is lower than that of WOA. However, the adaptability of the WGWOA algorithm converges to a lower value when the number of iterations is low. This condition is because the WGWOA algorithm introduces the position updating method of WOA algorithm based on the GWOA algorithm, which has the convergence performance and the ability of global optimization. This finding proves the feasibility of the WGWOA algorithm to optimize VMD.

Therefore, the WGWOA algorithm mentioned in this paper is used to optimize the parameters K and σ of VMD and label various fault types to facilitate later training of the fault diagnosis model. The optimal parameter K and σ solutions for different fault types were obtained, as shown in Table 5.

Table 5. Optimal parameter K and σ solutions for different fault types.

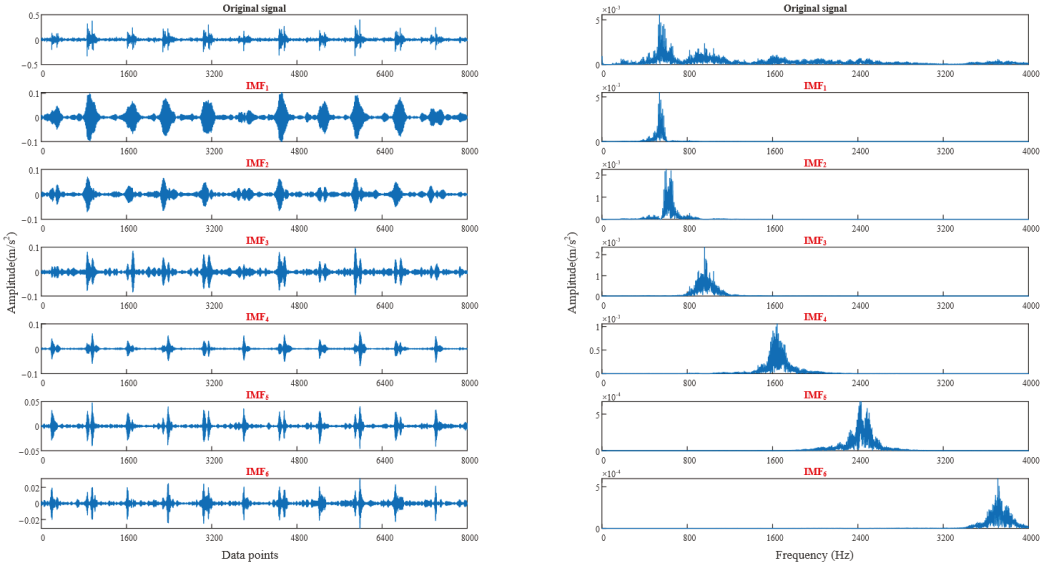
Fault Types	Optimum Solutions		Labels
	σ	K	
Normal	4835	6	1
Inner ring crack	4862	6	2
Outer ring crack	4822	6	3
Roller crack	4798	6	4

From Table 5, the optimum decomposition levels K of VMD for four fault type signals are all 6, optimized σ values are all around 4800. The method in Section 4.2 is adopted, the average value of σ is 4829.25 to form the optimal parameter combination [4829.25, 6]. The rolling bearing data of different fault types are decomposed through VMD by using the optimal parameter combination. The time-domain waveform and frequency spectrum of vibration signal after optimized VMD are shown in Figure 14. Only the vibration signal decomposition of the normal bearing and the bearing with damaged inner ring is listed here due to the length of the article.



(**a**)

Figure 14. *Cont.*



(**b**)

Figure 14. Optimizing VMD to decompose the time-domain waveform and spectrum of vibration signals of different fault types, (**a**) normal, (**b**) inner ring crack.

As shown in Figure 14, the IMF components of the two fault signals after VMD do not produce modal aliasing, which further verifies the feasibility of the WGWOA algorithm to optimize VMD.

5.2.3. Feature Extraction

In accordance with the proposed method of extracting eigenvalues, this paper extracts the permutation entropy of each component decomposed by VMD as the eigenvector F_v, so as to construct the data of SVM training samples and test samples. The expression of F_v is as follows:

$$F_v = [PE(\mathrm{IMF}_1), PE(\mathrm{IMF}_2), PE(\mathrm{IMF}_3), PE(\mathrm{IMF}_4), PE(\mathrm{IMF}_5), PE(\mathrm{IMF}_6)]^{\mathrm{T}} \quad (25)$$

The resulting data set is shown in Table 6.

Although a certain correlation is found between the permutation entropy of each component and the fault type, the specific relationship between them is more complex, and visually observing what fault type the eigenvector represents is impossible.

Table 6. Permutation entropy eigenvalue of vibration signal extraction.

Fault Types	Permutation Entropy					
	IMF_1	IMF_2	IMF_3	IMF_4	IMF_5	IMF_6
Normal	1.5309	1.0415	1.5753	0.8844	1.3019	0.1141
	1.3952	1.1453	1.7707	1.2923	1.3228	0.1434
	1.4194	1.0538	1.8166	0.9955	1.1806	0.1031
	1.2777	1.0539	1.8729	0.9517	1.0573	0.1127
	1.3725	1.1848	1.9700	1.0411	1.3821	0.1062
Inner ring crack	1.4377	1.6552	1.5502	0.8704	0.8514	0.1254
	1.4575	1.3057	2.0078	0.7103	0.8670	0.1349
	1.3202	1.5059	1.8016	0.8518	1.0440	0.1048
	1.2304	1.4806	1.8627	0.9084	1.0683	0.1600
	1.4751	1.4448	1.5748	0.7900	0.8558	0.1218
Outer ring crack	2.4565	2.2353	2.2428	1.5334	2.6846	0.3680
	1.7272	2.5163	2.0670	1.7688	2.7149	0.3893
	1.7458	2.4905	2.2135	1.5629	2.6006	0.4480
	1.7611	2.4239	2.1152	1.4379	2.5880	0.4388
	1.7432	2.4397	2.2388	1.5594	2.6276	0.4215
Roller crack	0.8964	1.4693	1.9504	1.3214	1.2628	0.2148
	1.2302	1.1797	2.1101	0.9923	1.7222	0.3233
	1.1171	1.3704	1.9537	0.8106	1.4210	0.2868
	1.2216	0.8641	2.0169	0.9422	1.4858	0.2865
	1.3158	1.1262	1.8119	0.9066	1.5658	0.2998

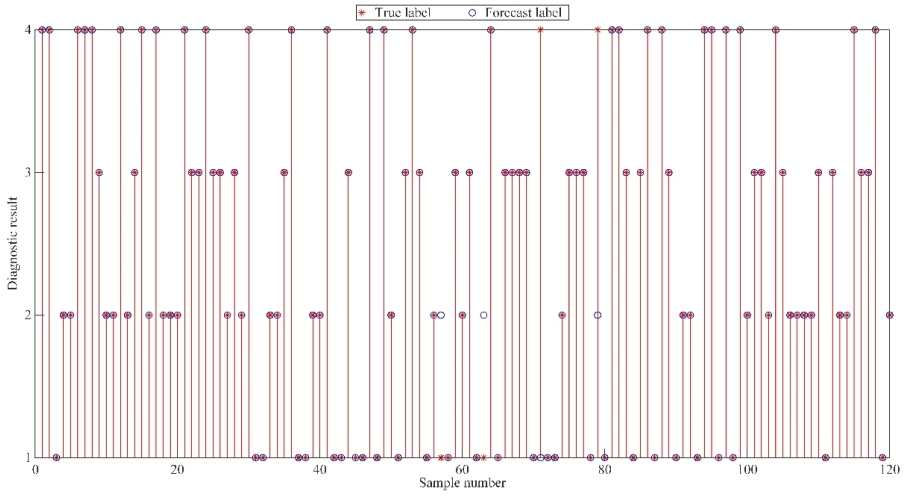
5.3. Fault Diagnosis Based on WGWOA-Optimized SVM

This paper uses SVM as the fault diagnosis model due to its powerful ability to process complex data. The proposed WGWOA algorithm is utilized to optimize its parameters c and g. The vibration signal eigenvector is processed in accordance with the method in Section 5.2.1 to construct the training and test samples, so as to train the SVM. The classification results of training samples and the diagnostic results of test samples are shown in Figure 15 and Table 7.

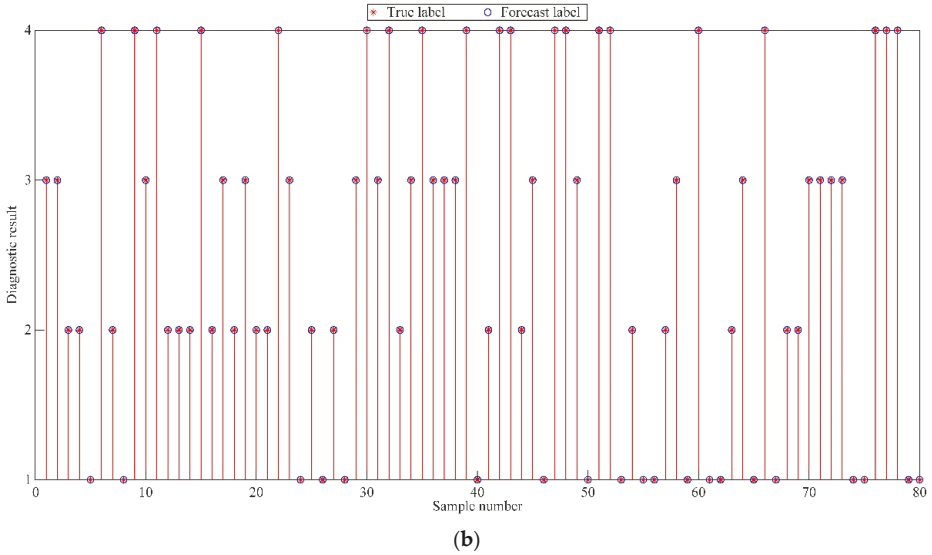
Table 7. Diagnostic error types of different samples.

Sample Types	Sample Point Label of Diagnostic Error	Actual Fault Types	Diagnostic Fault Types	Diagnostic Accuracy
Training sample	57	Normal	Inner ring crack	96.67%
	63	Normal	Inner ring crack	
	71	Roller crack	Normal	
	79	Roller crack	Inner ring crack	
Test sample	-	-	-	100.00%

As shown in Figure 15 and Table 7, only four sample points failed to be correctly classified in the SVM training process, the diagnostic accuracy of the training samples reached 96.67%, and the classification accuracy of the test samples reached 100%. Combined with the diagnostic accuracy of the two samples, the optimized SVM did not exhibit fitting phenomenon. The proposed optimization method of the permutation entropy characteristic matrix of each mode after VMD is scientific and effective because VMD can effectively avoid the phenomenon of signal mode aliasing, and the decomposed multiple modes are distinguished. The WGWOA algorithm is used to find the best K and σ parameter combination, which enhances the availability of VMD to extract the permutation entropy feature vector. The reliability of applying the WGWOA algorithm to the optimization of SVM parameters c and g is verified. This finding is because the WGWOA algorithm can efficiently and accurately find the optimal parameters c and g of SVM and build a high-performance SVM model to avoid over fitting and over learning.



(**a**)



(**b**)

Figure 15. Diagnostic results of different samples, (**a**) training samples, (**b**) test samples.

5.4. Comparative Analysis with Other Methods

The proposed fault diagnosis method was compared with the fault diagnosis methods of BPNN, SVM, EMD-SVM, VMD-SVM, WOA simultaneously optimizing VMD, SVM model (WOA-VMD-SVM), GWO simultaneously optimizing VMD, and SVM model (GWO-VMD-SVM) to verify its effectiveness and practicability. The above seven fault diagnosis methods were used for five experiments to increase the reliability of the experimental results and to avoid their randomness. The diagnostic results are shown in Table 8 and Figure 16.

Table 8. Diagnostic accuracy of different method tests.

Methods	Accuracy (%)					
	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Average
BPNN	72.50	61.25	63.75	52.50	76.25	65.25
SVM	76.25	76.25	72.50	80.00	76.25	76.25
EMD-SVM	80.00	82.50	76.25	73.75	81.25	78.75
VMD-SVM	87.50	87.50	90.00	81.25	85.00	86.25
WOA-VMD-SVM	96.25	92.50	93.75	95.00	93.75	94.25
GWO-VMD-SVM	96.25	96.25	98.75	98.75	92.50	96.50
WGWOA-VMD-SVM	100.00	100.00	98.75	100.00	100.00	99.75

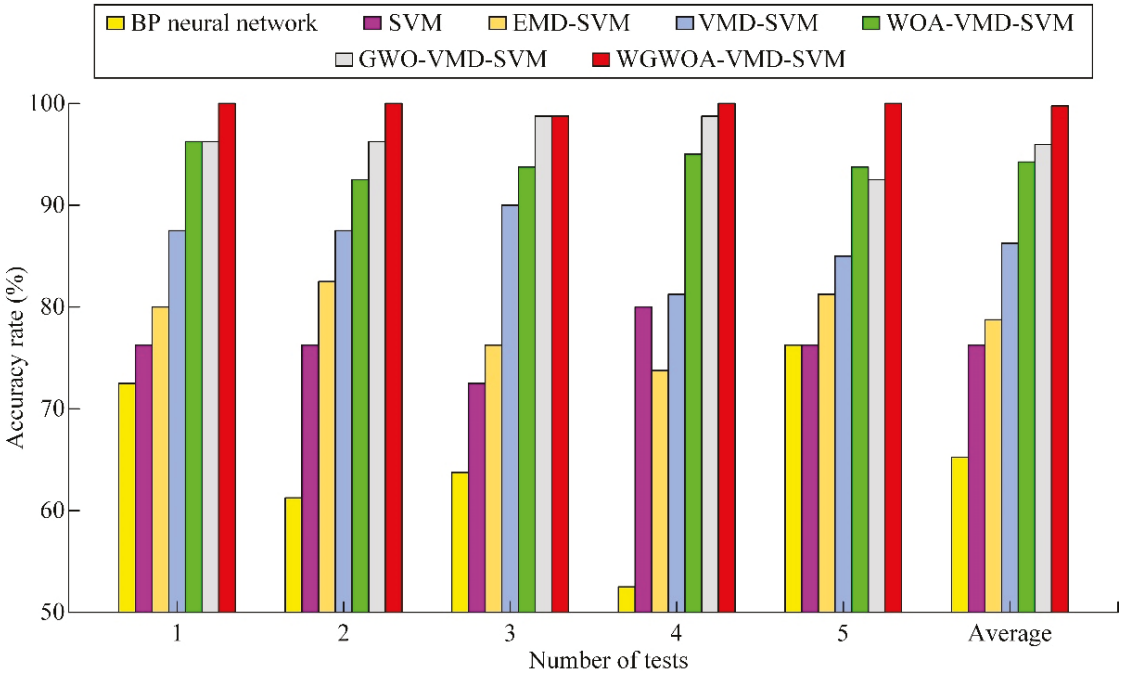


Figure 16. Test diagnostic accuracy.

As shown in Table 8 and Figure 16, the convergence speed of BPNN is slow, and the network performance is biased compared with SVM. The generalization ability of the network is poor under the small sample data, resulting in a draw accuracy of only 65.25%. This finding proves that the SVM fault diagnosis model has strong robustness under the small sample data. The average fault diagnosis rate of EMD-SVM model is 78.75%, which is higher than that of the SVM model. However, taking the normal bearing as an example, the peaks in the IMF component of its vibration signal decomposed by EMD appear at about 500 Hz and 1000 Hz in the IMF_2 spectrum, and modal aliasing occurs. EMD decomposes 12 IMF components (the last one is residual), and some component signals are arranged disorderly. This finding shows that decomposing noncharacteristic false components is extremely possible and extracting eigenvalues from the decomposed false components certainly increases the recognition difficulty of the fault diagnosis model. The average fault diagnosis rate of the VMD-SVM model reaches 86.25%, which is 10% higher than that of SVM. Combined with Figures 14 and 17, VMD has superior performance, and EMD is more suitable for actual fault diagnosis.

Although the VMD-SVM model is better than the BPNN, SVM, and EMD-SVM, it does not scientifically select the parameters of VMD and SVM, resulting in a fault diagnosis rate of less than 90.00%, and the model performance still needs to be improved. The average fault diagnosis of the model reaches 94.25% after optimizing the parameters of VMD and SVM with the WOA algorithm. This finding indicates that the WOA algorithm can play a certain role in the parameter optimization of VMD and SVM. However, the best c and g solutions obtained by WOA are 4.23 and 0.01 by observing Figure 18 and Table 9. Taking the cross-validation during the training of SVM as the fitness, the best and average fitness curves of WOA algorithm are maintained at a low level, and the best fitness convergence value is 91.67. For the other two algorithms, the SVM parameters obtained by the WOA algorithm may be local optimum. The best c and g solutions found by GWO are 15.32 and 0.22. Compared with WOA, the GWO algorithm has the best and higher average fitness curve, but it converges 28 times during iteration, which shows that the convergence of GWO algorithm is slower than that of the WOA algorithm, which is the same as that in Section 5.2.2. The WGWOA algorithm only converges to 96.67 at the best fitness of five generations. Compared with the WOA and GWO algorithms, the best and average fitness of the WGWOA algorithm are maintained at a high level. The average diagnostic rate of the WGWOA-VMD-SVM model for five repeated tests is 99.75%, which verifies the superiority of the WGWOA algorithm in SVM optimization. In conclusion, the proposed WGWOA-VMD-SVM method has many advantages, such as high efficiency and high accuracy, to meet the practical application requirements.

Table 9. Optimal solution of SVM parameters found by different algorithms.

Optimization Algorithms	Optimal Solutions	
	c	g
WOA	4.23	0.01
GWO	15.32	0.22
WGWOA	25.78	2.48

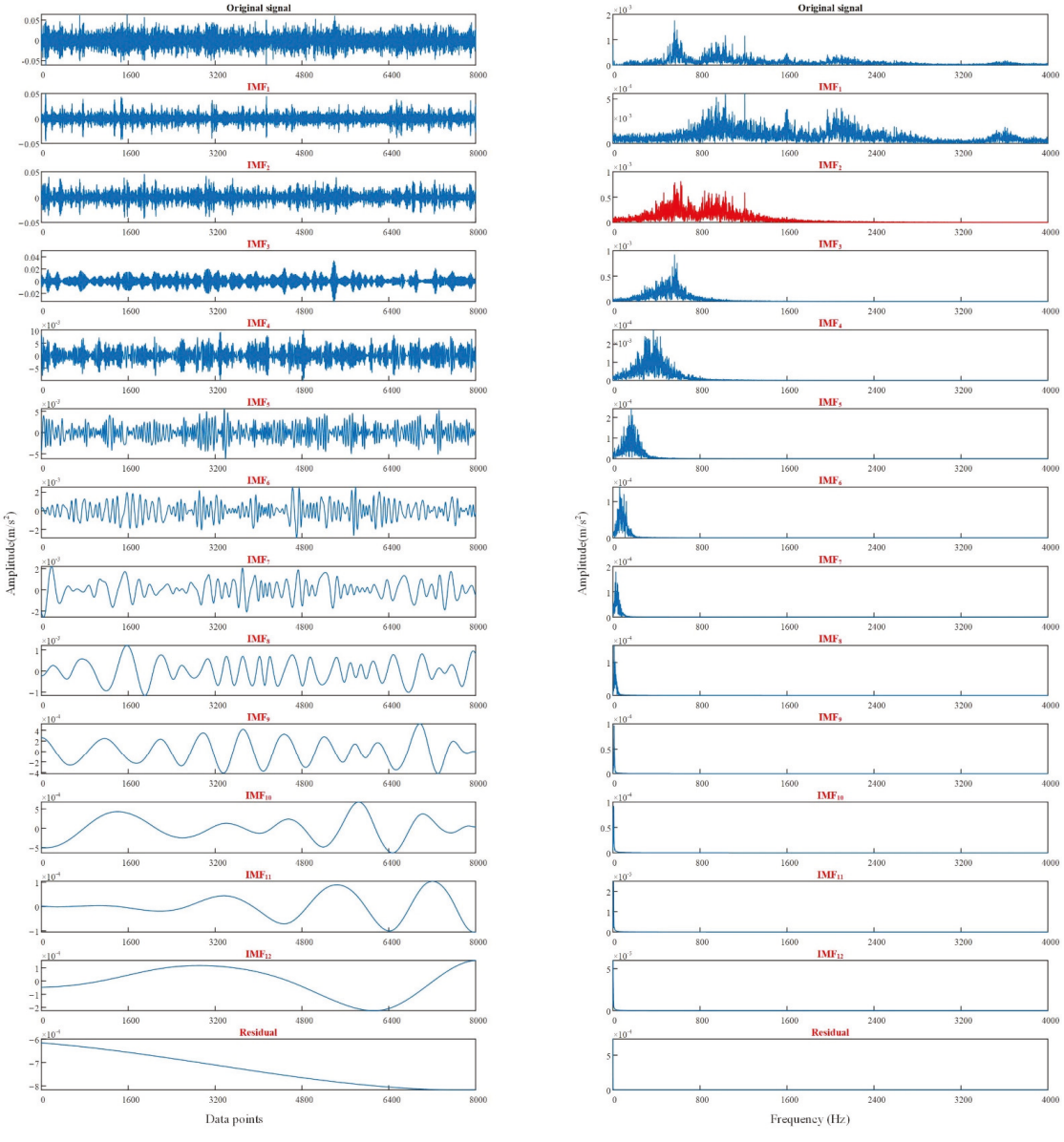


Figure 17. EMD result of normal bearing vibration signal.

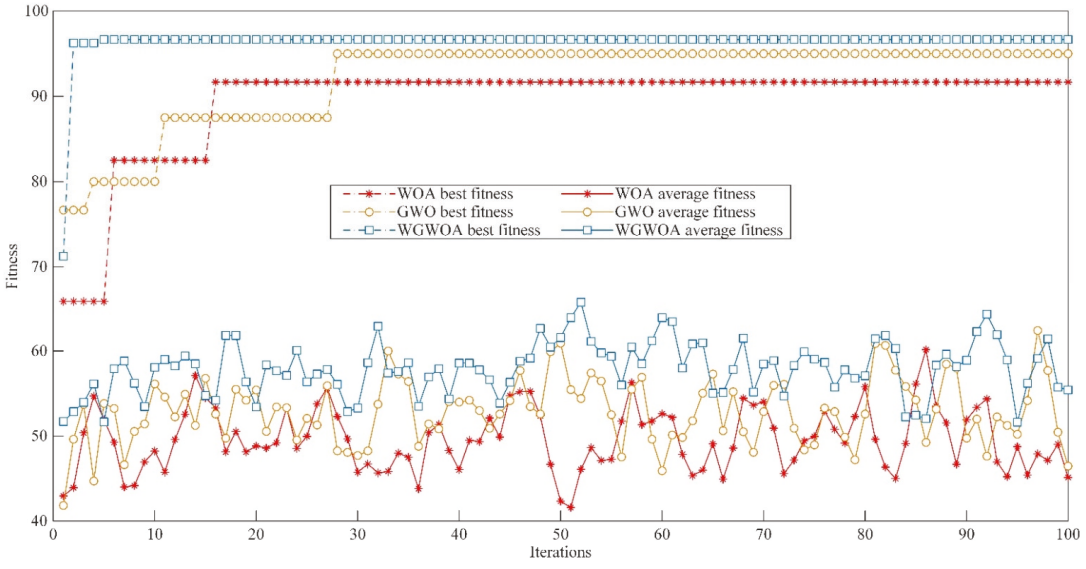


Figure 18. Different algorithms used to optimize the fitness curve of SVM.

6. Conclusions

In this paper, a fault diagnosis method combining VMD and SVM is adopted, and a WGWOA algorithm is proposed to optimize the K and σ parameters of VMD and the c and g parameters of SVM. The permutation entropy feature matrix is extracted, and the SVM is trained and verified by collecting the vibration signals of rolling bearings with different fault types for preprocessing. The conclusions after comparing the proposed method with several existing fault diagnosis methods are as follows:

1. The test results of two cases show that WGWOA-optimized VMD can properly suppress modal aliasing and that WGWOA-optimized SVM enhances the accuracy and self-adaptability of model classification. The average accuracy of this method in five repeated tests were 100.00% and 99.75%. Compared with other existing fault diagnosis methods, this method has many advantages, such as high accuracy and stable performance, to provide an effective new method for the existing fault diagnosis technology;
2. Compared with other optimization algorithms, the proposed WGWOA algorithm has good performance in terms of optimization accuracy, optimization efficiency, and algorithm convergence. The training process of this method is simple and fast, and the diagnostic accuracy after training is significantly higher than other traditional methods.

Author Contributions: Conceptualization, J.Z. and M.X.; methodology, J.Z.; software, J.Z. and Y.N.; validation, J.Z. and M.X.; formal analysis, M.X. and Y.N.; investigation, M.X. and J.Z.; resources, M.X. and G.J.; data curation, J.Z. and Y.N.; writing—original draft preparation, J.Z. and Y.N.; writing—review and editing, Y.N. and G.J.; visualization, M.X. and J.Z.; supervision, M.X. and Y.N.; project administration, M.X. and G.J.; funding acquisition, M.X. and G.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Jiangsu International Science and Technology Cooperation Project (No. BZ2022002), in part by the Agricultural Science and Technology Independent Innovation Fund of Jiangsu Province (No. CX(22)3101) and in part by the Key Research and Development Program of Jiangsu Province (No. BE2022385).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, C.B.; Li, H.K.; Zhao, X.W. Weak characteristic determination for blade crack of centrifugal compressors based on underdetermined blind source separation. *Measurement* **2018**, *128*, 545–557. [CrossRef]
2. Zhang, X.; Miao, Q.; Liu, Z.W.; He, Z.J. An adaptive stochastic resonance method based on grey wolf optimizer algorithm and its application to machinery fault diagnosis. *ISA Trans.* **2019**, *71*, 206–214. [CrossRef] [PubMed]
3. Wang, G.; Xiang, J.W. Remain useful life prediction of rolling bearings based on exponential model optimized by gradient method. *Measurement* **2021**, *176*, 109161. [CrossRef]
4. Islam, M.; Prosvirin, A.; Kim, J. Data-driven prognostic scheme for rolling-element bearings using a new health index and variants of least-square support vector machines. *Mech. Syst. Signal Process.* **2021**, *160*, 107853. [CrossRef]
5. Li, P.N.; Lei, Y.; Lin, J.S.; Ding, X. An improved exponential model for predicting remaining useful life of rolling element bearings. *IEEE Trans. Ind. Electron.* **2015**, *62*, 7762–7773. [CrossRef]
6. Wang, Z.Y.; Yao, L.G.; Chen, G.; Ding, J.X. Modified multiscale weighted permutation entropy and optimized support vector machine method for rolling bearing fault diagnosis with complex signals. *ISA Trans.* **2021**, *114*, 470–484. [CrossRef]
7. Yang, J.S.; Peng, Y.Z.; Xie, J.S.; Wang, P.X. Remaining Useful Life Prediction Method for Bearings Based on LSTM with Uncertainty Quantification. *Sensors* **2022**, *22*, 4549. [CrossRef]
8. Zhao, J.; Xiao, M.H.; Bartos, P.; Bohata, A. Dynamic engagement characteristics of wet clutch based on hydro-mechanical continuously variable transmission. *J. Cent. South Univ.* **2021**, *28*, 1377–1389. [CrossRef]
9. Xiao, C.A.; Tang, H.S.; Ren, Y.; Kumar, A. Fuzzy entropy assisted singular spectrum decomposition to detect bearing faults in axial piston pump. *Alex. Eng. J.* **2022**, *61*, 5869–5885. [CrossRef]
10. Chen, W.; Li, J.N.; Wang, Q.; Han, K. Fault feature extraction and diagnosis of rolling bearings based on wavelet thresholding denoising with CEEMDAN energy entropy and PSO-LSSVM. *Measurement* **2021**, *172*, 108901. [CrossRef]
11. Xiao, M.H.; Liao, Y.B.; Bartos, P.; Filip, M.; Geng, G.S.; Jiang, Z.W. Fault diagnosis of rolling bearing based on back propagation neural network optimized by cuckoo search algorithm. *Multimed. Tools Appl.* **2022**, *81*, 1567–1587. [CrossRef]
12. Chen, D.N.; Zhang, Y.D.; Yao, C.Y.; Sun, F.; Zhou, N.Y. Fault diagnosis based on FVMD multi-scale permutation entropy and GK fuzzy clustering. *J. Mech. Eng.* **2018**, *54*, 16–27. [CrossRef]
13. Peng, Z.; Chu, F.; He, Y. Vibration signal analysis and feature extraction based on reassigned wavelet scalogram. *J. Sound Vib.* **2002**, *253*, 1087–1100. [CrossRef]
14. Cheng, C.; Zhou, B.T.; Ma, G.J.; Wu, D.R.; Yuan, Y. Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data. *Neurocomputing* **2020**, *409*, 35–45. [CrossRef]
15. Zuo, L.; Xu, F.J.; Zhang, C.H.; Xiahou, T.F.; Liu, Y. A multi-layer spiking neural network-based approach to bearing fault diagnosis. *Reliab. Eng. Syst. Saf.* **2022**, *225*, 108561. [CrossRef]
16. Sun, Y.J.; Li, S.H.; Wang, X.H. Bearing fault diagnosis based on EMD and improved Chebyshev distance in SDP image. *Measurement* **2021**, *176*, 109100. [CrossRef]
17. Zhang, Y.T.; Li, C.L.; Jiang, Y.Q.; Sun, L.; Zhao, R.B.; Yan, K.F.; Wang, W.H. Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. *J. Clean. Prod.* **2022**, *354*, 131724. [CrossRef]
18. Deng, Y.; Zhu, K.H.; Zhao, G.J.; Zhu, J.Y. Efficient partial discharge signal denoising method via adaptive variational modal decomposition for infrared detectors. *Infrared Phys. Technol.* **2022**, *125*, 104230. [CrossRef]
19. Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [CrossRef]
20. Guo, Y.; Yang, Y.; Jiang, S.; Jin, X.; Wei, Y. Rolling Bearing Fault Diagnosis Based on Successive Variational Mode Decomposition and the EP Index. *Sensors* **2022**, *22*, 3889. [CrossRef]
21. Lin, Y.; Xiao, M.H.; Liu, H.J.; Li, Z.L.; Zhou, S.; Xu, X.M.; Wang, D.C. Gear fault diagnosis based on CS-improved variational mode decomposition and probabilistic neural network. *Measurement* **2022**, *192*, 110913. [CrossRef]
22. Ayman, Y.; Mohammed, E.; Abdelhalim, Z. The role of artificial intelligence in photovoltaic systems design and control: A review. *Renew. Sustain. Energy Rev.* **2017**, *78*, 72–79.
23. Wei, W.; Cong, R.; Li, Y.T.; Abraham, A.; Yang, C.Y.; Chen, Z.T. Prediction of tool wear based on GA–BP neural network. *Proc. Inst. Mech. Eng.* **2022**, *236*, 1564–1573. [CrossRef]
24. Wei, W.; Shang, Y.; Peng, Y.; Cong, R. Prediction Model of Sound Signal in High-Speed Milling of Wood–Plastic Composites. *Materials* **2022**, *15*, 3838. [CrossRef]
25. Xiao, M.H.; Zhang, W.; Wen, K.; Zhu, Y.; Yiliyasi, Y. Fault diagnosis based on BP neural network optimized by beetle algorithm. *Chin. J. Mech. Eng.* **2021**, *34*, 119. [CrossRef]
26. Mahdi, M.; Mahdieh, G.; Hossein, M. A hybrid intelligent approach to detect Android Botnet using Smart Self-Adaptive Learning-based PSO-SVM. *Knowl. Based Syst.* **2021**, *222*, 106988.

27. Wumaier, T.; Xu, C.; Guo, H.Y.; Jin, Z.J.; Zhou, H.J. Fault Diagnosis of Wind Turbines Based on a Support Vector Machine Optimized by the Sparrow Search Algorithm. *IEEE Access* **2021**, *9*, 69307–69315.
28. Chen, H.; Li, S. Multi-Sensor Fusion by CWT-PARAFAC-IPSO-SVM for Intelligent Mechanical Fault Diagnosis. *Sensors* **2022**, *22*, 3647. [CrossRef]
29. Van, M.; Kang, H.J. Bearing defect classification based on individual wavelet local fisher discriminant analysis with particle swarm optimization. *IEEE Trans. Ind. Inform.* **2015**, *12*, 124–135. [CrossRef]
30. Hou, Y.; Gao, H.; Wang, Z.; Du, C. Improved Grey Wolf Optimization Algorithm and Application. *Sensors* **2022**, *22*, 3810. [CrossRef]
31. Chen, B.; Zhou, C.; Liu, Y.; Liu, J.H. Correlation analysis of runway icing parameters and improved PSO-LSSVM icing prediction. *Cold Reg. Sci. Technol.* **2022**, *193*, 103415. [CrossRef]
32. García Nietoa, P.J.; García-Gonzaloa, E.; Sánchez Lasherasb, F.; De Cos Juezc, F.J. Hybrid PSO-SVM-based method for forecasting of the remaining useful life for aircraft engines and evaluation of its reliability. *Reliab. Eng. Syst. Saf.* **2015**, *138*, 219–231. [CrossRef]
33. Dong, Z.L.; Zheng, J.D.; Huang, S.Q.; Pan, H.Y.; Liu, Q.Y. Time-shift multi-scale weighted permutation entropy and GWO-SVM based fault diagnosis approach for rolling bearing. *Entropy* **2019**, *21*, 621. [CrossRef] [PubMed]
34. Zheng, H.B.; Zhang, Y.Y.; Liu, J.F.; Wei, H.; Zhao, J.H.; Liao, R.J. A novel model based on wavelet LS-SVM integrated improved PSO algorithm for forecasting of dissolved gas contents in power transformers. *Electr. Power Syst. Res.* **2018**, *155*, 196–205. [CrossRef]
35. He, B.H.; Jia, B.Y.; Zhao, Y.H.; Wang, X.; Wei, M.; Dietzel, R. Estimate soil moisture of maize by combining support vector machine and chaotic whale optimization algorithm. *Agric. Water Manag.* **2022**, *267*, 107618. [CrossRef]
36. Mirjalili, S.; Mirjalili, S.M.; Lewis, A. Grey Wolf Optimizer. *Adv. Eng. Softw.* **2014**, *69*, 46–61. [CrossRef]
37. Mehdi, G.; Reza, J.Y.; Amir, H.D. Building energy optimization using Grey Wolf Optimizer (GWO). *Case Stud. Therm. Eng.* **2021**, *27*, 101250.
38. Song, X.H.; Tang, L.; Zhao, S.T.; Zhang, X.Q.; Li, L.; Huang, J.Q.; Cai, W. Grey Wolf Optimizer for parameter estimation in surface waves. *Soil Dyn. Earthq. Eng.* **2015**, *75*, 147–157. [CrossRef]
39. Mirjalili, S.; Lewis, A. The Whale Optimization Algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]
40. Rajabi, S.; Azari, M.S.; Santini, S.; Flammini, F. Fault diagnosis in industrial rotating equipment based on permutation entropy, signal processing and multi-output neuro-fuzzy classifier. *Expert Syst. Appl.* **2022**, *206*, 117754. [CrossRef]
41. Zhang, X.Y.; Li, C.S.; Wang, X.B.; Wu, H.M. A novel fault diagnosis procedure based on improved symplectic geometry mode decomposition and optimized SVM. *Measurement* **2021**, *173*, 108644. [CrossRef]
42. Zhou, J.M.; Wang, F.L.; Zhang, C.C.; Zhang, L.; Yin, W.H.; Li, P. An intelligent method for rolling bearing evaluation using feature optimization and GA-SVM. *J. Vib. Shock* **2021**, *40*, 227–234.
43. Li, Y.H. *Research on Dynamic Characteristics Modeling and Simulation for a Faulty Ball Bearing under Variable Speed and Load*; East China Jiaotong University: Nanchang, China, 2006.

sensors

MDPI

Article

Miniterm, a Novel Virtual Sensor for Predictive Maintenance for the Industry 4.0 Era

Eduardo Garcia [1,†], Nicolás Montés [2,†,*], Javier Llopis [1] and Antonio Lacasa [1]

[1] Ford Spain, Poligono Industrial Ford S/N, 46440 Valencia, Almussafes, Spain
[2] Department of Mathematics, Physics and Technological Sciences, University CEU Cardenal Herrera, C/ San Bartolome 55, 46115 Valencia, Alfara del Patriarca, Spain
* Correspondence: nicolas.montes@uchceu.es
† These authors contributed equally to this work.

Abstract: This article introduces a novel virtual sensor for predictive maintenance called mini-term. A mini-term can be defined as the time it takes for a part of the machine to do its job. Being a technical sub-cycle time, its function has been linked to production. However, when a machine or component gets deteriorated, the mini-term also suffers deterioration, allowing it to be a multifunctional indicator for the prediction of machine failures as well as measurement of production. Currently, in Industry 4.0, one of the handicaps is Big Data and Data Analysis. However, in the case of predictive maintenance, the need to install sensors in the machines means that when the proposed scientific solutions reach the industry, they cannot be carried out massively due to the high cost this entails. The advantage introduced by the mini-term is that it can be implemented in an easy and simple way in pre-installed systems since you only need to program a timer in the PLC or PC that controls the line/machine in the production line, allowing, according to the authors' knowledge, to build industrial Big Data on predictive maintenance for the first time, which is called Miniterm 4.0. This article shows evidence of the important improvements generated by the use of Miniterm 4.0 in a factory. At the end of the paper we show the evolution of TAV (Technical availability), Mean Time To Repair (MTTR), EM (Number of Work order (Emergency Orders/line Stop)) and OM (Labour hours in EM) showing a very important improvement as the number of mini-terms was increased and the Miniterm 4.0 system became more reliable. In particular, TAV is increased by 15%, OM is reduced in 5000 orders, MTTR is reduced in 2 h and there are produced 3000 orders less than when mini-terms did not exist. At the end of the article we discuss the benefits and limitations of the mini-terms and we show the conclusions and future works.

Keywords: Miniterm; IIoT; Industry 4.0; fault detection; sub-cicle time; virtual sensor

Citation: Garcia, E.; Montés, N.; Llopis, J.; Lacasa, A. Miniterm, a Novel Virtual Sensor for Predictive Maintenance for the Industry 4.0 Era. *Sensors* **2022**, *22*, 6222. https://doi.org/10.3390/s22166222

Academic Editors: Dong Wang, Shilong Sun and Changqing Shen

Received: 29 June 2022
Accepted: 15 August 2022
Published: 19 August 2022

1. Introduction

A production, manufacturing or assembly line can be defined as a group of sequential operations established in a factory where the product is moved through them while the final product is built. The design of these types of lines is the first key factor, there are a large number of crucial decisions that will be made during product design, line format configuration, line balancing, machine selection, technology available, etc. One of the most critical parameters when designing a line is the cycle time. This is defined as the time the part takes to be made. In a "perfectly" balanced line, it is the time spent on each position of the line to manufacture each of the parts for the final product. In general, these problems are considered one at a time, due to their complexity.In general, these problems are considered one at a time, due to their complexity. Hence, the great importance of the line balancing where the tasks are assigned to the work stations and taking into account the resources used. Due to the importance of this task, a large number of researchers have been working on this topic. Under the acronym ALB (Assembly Line Balancing), a large number of optimization models have been presented and discussed in scientific and technical literature [1].

1.1. What Happens after the Installation of the Production Lines?

Once the line has been installed and begins to produce parts, this comes under the control of the factory management. They need to find a technically feasible point in which maintenance and product quality levels should be selected to achieve the highest productivity in order to meet the company's objectives for greater profitability. The correct combination of three factors, production, quality and maintenance, is the key to competitiveness of the company.

Quality and production: The link between quality and production is perfectly defined in the literature [2,3]. The improvement of the quality implies eliminating, for example, imperfections that may cause rework, which would increase the production cost and, consequently, the final cost of the part.

Maintenance and production: Maintenance is a function of the factory management that runs in parallel with production. The main output of the production is the expected product and its secondary output is a maintenance demand, which is the input of the maintenance function. The result of maintenance is an additional production input considered as *production capacity*. While production manufactures the product, maintenance provides the production capacity. Thus, maintenance affects production by increasing its capacity and controlling the quantity and quality of the output (the product) [2].

Maintenance and quality: The role of maintenance in the long-term use of manufacturing lines is well known and taken into account by researchers and engineers. There are many strategies that seek to increase the long service life of the lines. However, extending the useful life of the line is pointless without quality criteria being maintained [2]. In general terms, the equipment that is not properly maintained has periodic failures, suffers speed losses and loses precision and therefore tends to cause faults [2]. On the other hand, excessive maintenance may result in unnecessary costs.

Maintenance, quality and production: Determining the optimal time in which to perform the maintenance task is the key to be able to link the three factors. In [4] an analytical study has been conducted for a food products company to see the link between these three factors. The link between maintenance and production is positive. This implies that, the more maintenance, the machines work in better conditions, generating a continuous production. The link between quality and production is negative. This implies that the quality control activities expose production. Lower production is obtained the more hours are invested in quality control.

1.1.1. Production

Focusing our attention on production, the line is designed by a team of experts, based on all the parameters, also defining the maximum capacity of production. In the automotive sector, this maximum capacity is measured by the JPH (jobs per hour) and it is known as "Engineering Running Capacity" (ERC). The objective of those responsible for production is to ensure that the line will reach this maximum value. The reality of the manufacturing lines shows that the ERC cannot be reached in practice, so the production engineers have estimated a new maximum production rate which is known as "Engineering Running Rate" (ERR). This new rate can be defined empirically, based on observations of the line performance. This mismatch between ERC and ERR can be due to a multitude of factors since, during the useful life of a production line, which can be decades, performance depends on a large number of parameters such as maintenance, stopping time events, equipment breakdowns, waiting systems, the dynamic behaviour of bottlenecks, the Bowl phenomenon, market demand, new available technologies, etc. In fact, each workstation has its own identity as they are not intrinsically identical [5]. This means that many of the simplified versions of the models or line balancing algorithms are not applicable to real lines, having to be adapted "manually" [5,6].

1.1.2. Maintenance

Another key factor in the performance of the manufacturing lines is maintenance. In general, maintenance can be classified as two main groups: Corrective Maintenance (CM) and Preventive/Predictive maintenance (PM). CM is carried out when the machine fails or some the equipment elements are damaged and must be replaced or repaired, this element and/or part will be responsible for a failure in the whole line if the action is not implemented. However, the PM is carried out before the equipment fails. The purpose of a PM order is to promote continuous production of the system and/or minimize the loss of performance. In the preventive/predictive maintenance, we can find two great types of strategies: based on time (Time-based Maintenance, TBM) or those based on the state of the machine (Condition Based Maintenance, CBM). Those strategies based on time involve carrying out a preventive maintenance periodically, lubricating, calibrating and performing periodic controls. Instead, the CBM strategy involves making a real-time diagnosis where the decision is made by observing the "condition" of the system and its components [6]. TBM strategies are based on the manufacturer's recommendations, fault history, operators and/or maintenance staff's experience. In contrast, in the CBM strategy, the objective is avoiding unnecessary maintenance tasks and performing them when there is evidence of abnormal functioning. It is a proactive strategy in which the development of a predictive model is required. The CBM is that 99 % of equipment failures are preceded by certain signs, conditions or indications that the failure is about to occur [7]. The system condition is quantified through sensor measurements taken periodically and even continuously [6,8]. In general, the purpose of the CBM is twofold. First, collecting data on the machine condition and second, increasing the knowledge of the causes for the failures, the effects and the deterioration patterns of the equipment [8]. In addition, CBM, through this strategy, can ensure a high quality of the final product, especially if the thresholds of the measurements being taken from the machine are correctly selected [2]. CBM can be carried out in two ways: online or offline [8]. The online process involves carrying it out while the machines are active. On the contrary, in offline mode, the process is performed while the machine is stopped. In this case we usually look for cracks, colour changes, etc. Moreover, the CBM can be done continuously or periodically. The most usual way is to do it periodically, for example, every hour or every change of shift, although the ideal way would be to do it continuously and automatically. However, as indicated in [9], it may be very expensive since many sensors and devices are needed to carry it out. The most commonly used sensors to perform the CBM are the following:

- *Vibration*: Vibration sensorization is one of the most commonly used techniques for CBM, especially for machines with rotating elements [10]. The analysis is done in situ and it is a non-destructive test.
- *Noise*: It is another of the most used techniques in CBM. It has a strong link with vibration and therefore it is also used for machines with rotating elements [10]. However, there is a fundamental difference between the two. While the sensorization of the vibration requires being in contact with the machine or element to be sensorized, noise monitoring is simply "listening" to the equipment without having to be in contact [8].
- *Analysis of the oil or lubricant*: With this technique, the oil is analysed to determine whether it is able to function or not properly. In addition, it also provides an indirect measure of the deterioration level of the components lubricated [8].
- *Electrical measurements*:This technique includes the measurement of changes in the properties of equipment such as resistance, conductivity and insulation. This technique is usually carried out to detect deterioration of insulation in engines,
- *Temperature*: This technique is used primarily to detect faults in electrical and electronic components [8].
- *Pressure, flow, electric consumption*: These techniques are also used, although to a lesser extent than the previous ones.

The decisions to be made under the CBM concept can be classified into two: Diagnosis and prediction. Diagnosis is the process of finding the source of the failure while prediction

is the process of estimating when the failure will occur [11]. The objective of the diagnosis is to warn maintenance engineers on equipment operations under abnormal functioning conditions. Even if the equipment is working in abnormal conditions, this does not mean the equipment has failed. This will happen after a certain time [8]. The time that remains until the failure happens is the one that must estimate the prediction. Regarding maintenance, prediction is much more relevant than the diagnosis since unexpected failures can be predicted [9].

1.1.3. Change Point Definition

At this point, what is known as "Change point" comes into play, see Figure 1.

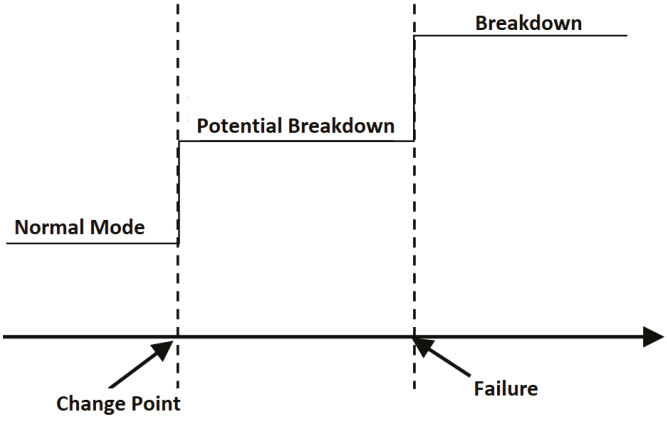


Figure 1. Change point definition .

The change point is defined as an abrupt change in the time series measurement that is being made of the machine, vibration, noise, etc. More formally, let us assume we have an ordered sequence of data, $y_{1:n} = (y_1, \ldots, y_n)$. A changepoint is said to occur within this set when there exists a time, $t \in \{1, \ldots, n-1\}$, such that the statistical properties of $\{y_1, \ldots, y_t\}$ and $\{y_{t+1}, \ldots, y_n\}$ are different in some way.

The change point is an indication that something anomalous is happening and announces the end of the useful life of some component. In [12] an attempt is made to define a guide on how to deal with CBM. As an example, Figure 2 shows the deterioration of an elevator measured with a vibration sensor at the Ford factory located in Almussafes (Valencia).

Figure 2. Change point of an elevator measured with a vibration sensor at the Ford factory located in Almussafes (Valencia).

The change point is always related with some physical change of the component. In the case of oil or lubricant, we know there is a sudden change in performance, mainly because, when the oil is approaching the end of its useful life, its viscosity changes abruptly. When a component or part is subjected to a constant load, the elongation that suffers over time is known as "the creep curve" where, at the end of its life there is an accelerated elongation, see [13]. Something similar happens with the elasticity coefficient. When a part is subjected to continuous bending, such as a train track, see [14,15], and the end of its useful life is approaching, there comes a point where it will not recover its starting position. There are different techniques to detect change points, such as EWMA, CUMSUM, MSE, etc., see [15]. Given their relevance to CBM, new techniques are being researched for more complex cases, see [16], and a special package has even been developed for R, see [17].

2. Previous Results by the Research Team

In [18] an improvement of the existing mathematical model in the literature on production lines was proposed, and its use in the improvement of the manufacturing process. The literature classifies the time data used in the analysis of the manufacturing process into two types, the long-term data (long-terms), and the short-term data (short-terms). Long-term data time are used mainly for process planning while short-term data time are used mainly for process control. There is abundant literature in which the analysis of long-term times is studied, in comparison with the literature that studies short-term times. Following the definition by [19], the short-term data refer to a time not long enough for the failure period of the machine and where the cycle time of the machine is considered short-term time.

2.1. From Long-Term to Micro-Term Cycle Time Data Model

In [18], the short-term is redefined in two new terms: the mini-term data time and the micro-term data time. A mini-term can be defined as the time it takes for a part of the machine to do its job. The division of a machine into mini-terms can be conditioned by, in a preventive maintenance policy or in a breakdown, the component that could be replaced in an easier and faster way than another subdivided part of the machine. A mini-term could also be defined as a sub-division that allows us to understand and study the machine behaviour. In the same way, a micro-term is defined as each part of the mini-term that could be divided into itself, see Figure 3. A mini-term can be calculated by adding the micro-terms into which the mini-term is divided. In the same way, a short-term can be calculated adding the mini-terms into which the short-term is divided. The same with long-term and short-term. For more information about this multi-scale time analysis in dynamical systems, please check our previous works [18,20,21].

The mathematical model proposed in [18] was reformulated in [20], using tensor algebra, which reduces the computational cost of the model, particularly when the number of mini-terms and micro-terms is high.

More recently, in [22] the data model is mixed with the complete modeling of a factory using Petri nets to develop a manufacturing map, a hierarchical construction of Petri nets in which the lowest level network is a temporary Petri net based on mini-terms, and in which the highest level is a global view of the entire plant. The user of a manufacturing map can select intermediate levels, such as a specific production line, and perform analysis or simulation using real-time data from the mini-term database.

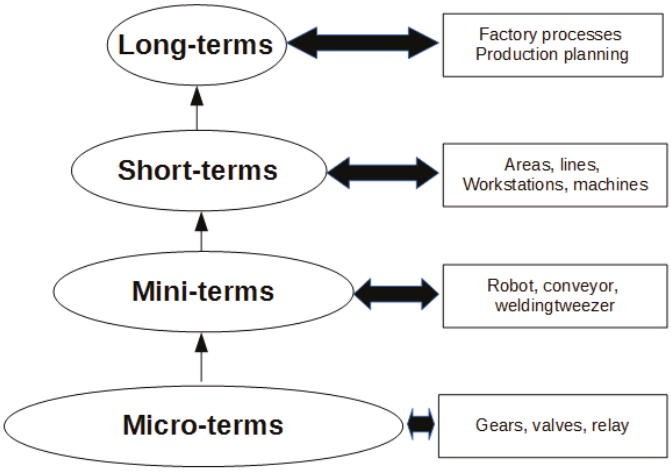


Figure 3. From the micro-term to the long-term.

2.2. The Mini-Term: The Link between Production and Maintenance

One of the contributions made in [18] was the use of the mini-term to link production and maintenance. The mini-term, by definition is a sub-cycle time and it would only make sense to use it for production improvement. More formally, the mini-term can be defined as an ordered sequence of sub-cycle times, $m_{1:n} = (m_1, \ldots, m_n)$.

However, in [18] it is shown that when a deterioration happens and ends up becoming a change point in vibration sensors (see Figure 2), this change point also happens in the mini-term, that is, a physical change point will result in a deterioration of the cycle time. More formally, a change point in the mini-term is said to occur within this set when there exists a time, $t \in \{1, \ldots, n-1\}$, such that the statistical properties of $\{m_1, \ldots, m_t\}$ and $\{y_{m+1}, \ldots, _n\}$ are different in some way. Figure 4 shows two examples of change points in mini-terms of a welding clamp measured at Ford factory in Almussafes. The first is due to the deterioration of the proportional valve that controls the clamp movement. The second is due to an internal leak in the clamp cylinder.

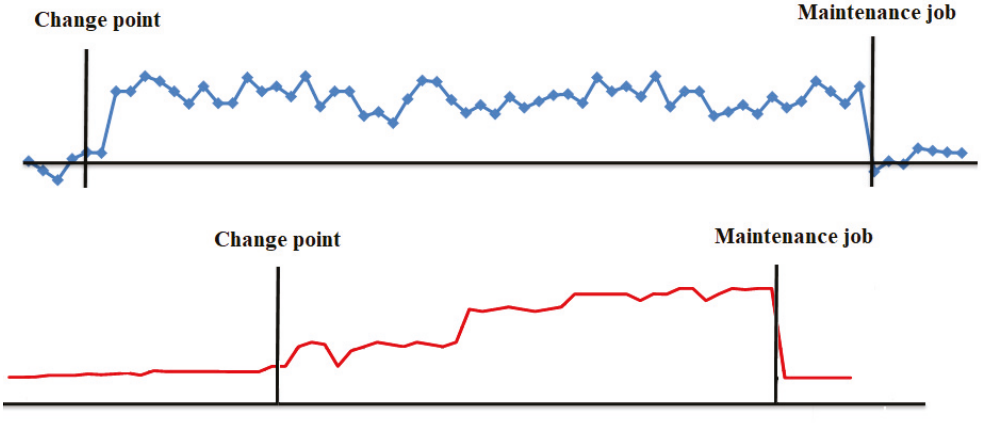


Figure 4. Change points in mini-terms. Proportional valve (**above**). Leak in cylinder (**below**).

In the mini-term we can see reflected all the physical change points that take place in the machine, and these change points result in a deterioration of the sub-cycle time. The example in Figure 5 shows the effect on the mini-term that the deterioration of the lubricant in the welding clamp has and how, once lubricated correctly, it recovers its nominal value. It means that mini-term colud act as a "virtual" sensor for predictive maintenance tasks.

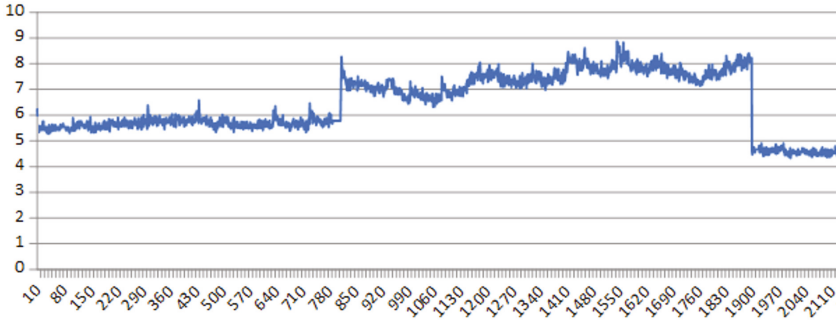


Figure 5. Change points in mini-terms of a welding clamp due to lack of lubrication.

2.3. Benefits Using Mini-Term in the Industry 4.0 Era

One of the main handicaps of Industry 4.0 is the cost of adding sensors in the machines and how to integrate them with the installed systems. As explained in the previous section, there are different prediction systems based on vibration, noise, temperature sensors, etc., but they are excessively expensive if we think about using them in a massive way for all the machines/components that we can find in a car factory, for example. It would take a large number of sensors, wiring the installation, programming the measurement, etc., and for this reason, these techniques are not used in a massive way, only for critical machines. Thus, the success or failure of the proposals related to IIOT (Industrial Internet of Things) and Industry 4.0 is mainly influenced by the cost of the proposals, the number of devices to connect and their interaction with the pre-installed systems in the production process.

However, mini-terms (technical sub-cycle times) could be measure with the PC/PLC sensors and the industrial network already installed in the production line which is responsible for the automatic production of the lines, doing the installation process, cheap and easy to install. It allows, for the first time, to create a IIOT Big Data based on Miniterms to predict failures in production lines

2.4. Installation Setup

As mentioned above, the objective is to measure the mini-terms (technical sub-cycle times) with the PC/PLC sensors and the industrial network already installed in the production line which is responsible for the automatic production of the lines. To do this, the first step was to include the mini-terms in Ford's standards in order to begin mass programming. In large companies like Ford, there are standards and protocols used to program PLCs, with I/O restrictions, memory, etc. No supplier can program a PLC if he/she does not know the standard.

Thanks to standardization, it is currently possible to program a mini-term in any PLC that Ford has in any factory in the world where the industrial network is in charge of channeling the mini-terms to the Database. With this it is possible to measure the mini-term, with the only cost of programming a timer in the PLC/PC allowing, for the first time, the massive monitoring of the time it takes for the components of the machines to carry out their task.

In order to activate and monitor the mini-terms and implement the project, a software interface programmed in R was developed, see Figure 6.

Figure 6. Architecture for Miniterm 4.0. It collects Miniterms in Real-time at Ford factories.

3. Objective of Our Line of Research

The objective of our line of research is to create the intelligent system "Miniterm 4.0" where, through real-time monitoring of the mini-terms, we will be able to predict the failures of the monitored components, determine their pathology and determine the effect they will have on production and the quality of the manufactured part, see Figure 7. This article shows the results for the first part of the goal: predict the failures. During the last years, Ford factory in Almussafes (Valencia) has been programming mini-terms in different components like cylinders, clamps, robots, elevators, conveyors, gearbox, fans, switches, turntables, etc. Nowadays, the number of mini-terms are up to 22 K. During the installation process, the availability indicators of the machines and production have improved significantly as the mini-terms have been installed and the detection algorithms have been improved. Section 4 shows the evolution and current algorithm for detecting change points by using mini-terms, the common causes that show false positives and how they are being solved. Section 5 shows the results and goodness of the system from the point of view of the improvement in the production indicators, TAV, OM EM, etc. Section 6 shows the conclusions and future work.

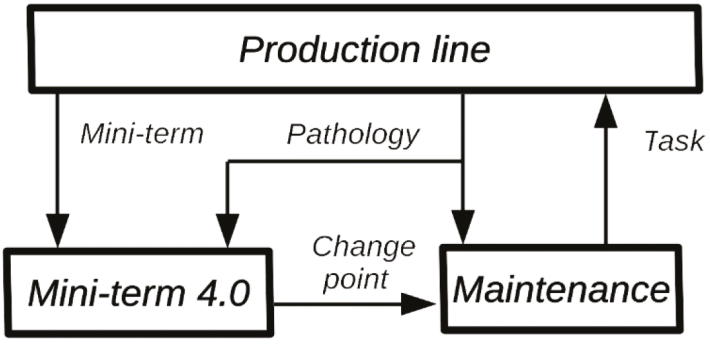


Figure 7. Miniterm 4.0 architecture to predict failures through mini-terms.

4. Towards Robust Detection of the "Change Point" of Mini-Term

The mini-terms installation process started in mid-2018. In the first version of the application, a basic state change detection algorithm was implemented to detect change points that would allow us to automatically detect when a mini-term had changed its behaviour significantly. The algorithm implemented was a K-Means classification algorithm [21]. This algorithm divided the data accumulated in the previous 9 days into two groups, based on the mean value of the groups. Two thresholds were established. A 7% and an 18%, the first one was used to set up an alarm configuring that component in warning mode and the second to automatically send an e-mail to the maintenance workers so that they could proceed to check the component [21]. This algorithm allowed us to store time series with change points that took place in the plant. After several months accumulating in the database the real cases detected, we were able to get an idea of the different types of change points that could exist. The 10 most representative ones were selected and all the change point algorithms programmed in R were tested [23]. A comparison was made, both in its effectiveness and precision in detecting the change point, as well as in its computation time. From this comparison, the result showed that Bartlett's algorithm was the most effective one in which the version of the winding sliding algorithm is the most efficient at computational level [24]. This algorithm was implemented in the system, replacing the previous one. This new version of the algorithm made it possible to replace the change point based on a percentage of the mean by a *p*-value, the percentage of success showing that a change point has actually taken place, as well as its location in the time series. These two improvements led to an improvement in detection, but there were also new false positives and negatives caused by:

- ***Oscillatory change point***: There are pathologies/components that, when an anomaly occurs, this results in a fluctuation in the mini-term but it does not result in a change in the mean value of the data, making it undetectable by Bartlett's algorithm, causing false negatives.
- ***Slow deterioration***: There are pathologies that generate a slow deterioration where the mini-term increases its value progressively but slowly and it is necessary to compare it with previous months to determine the deterioration caused. False negatives.
- ***Scan-Time (PLC's sampling frequency)***: One of the handicaps that have come up due to the massive use of mini-terms is the Scan-Time. Scan-Time is the time it takes for the PLC to collect the inputs, execute its PLC program and then update the outputs. Since the objective is to use the PLC's already installed to measure the mini-terms, it is a parameter that is imposed and generates false positives. Figure 8 shows the effect produced by the Scan-Time on the measured data. This effect causes false positives when the Scan-Time is high in relation to the mini-term value.

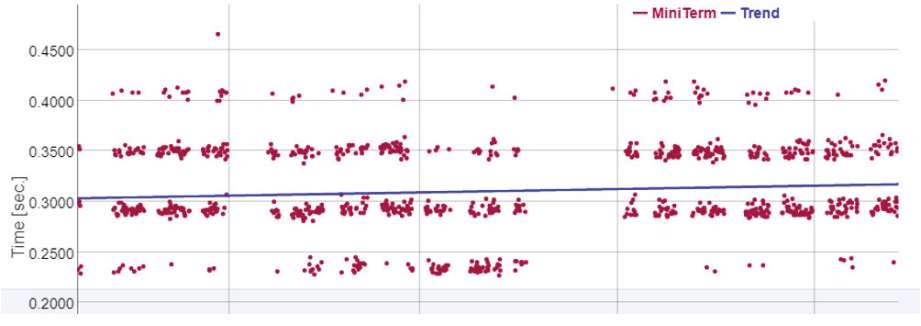


Figure 8. Change point produced by Scan-Time.

5. Current Mini-Term Anomaly Detection Algorithm

In the literature, an initial calibration is usually carried out to adjust the algorithms to the initial state of the machines or components, see [25]. On the other hand, there is a widespread use of numerical indicators in the detection of change points in the industry, see [26–28]. The great advantage of using numerical indicators is the possibility of adapting and developing new indicators in order to adapt them to different types of characteristics. However, this advantage could be a problem if you want to generalize the indicators for whatever component, that is the case of mini-terms. Therefore, a robust and general algorithm is welcome for those cases.

Most common indicators, see [26–28], Mean, variance, Skewness, Kurtosis, etc., have the same initial supposition, that is, the data you collect for a machine with good health follows a normal distribution. In statistics, when you have a set of data that follows a normal distribution, the methods to detect outliers are well known. In statistics, an outlier is a data point that differs significantly from other observations.These values are commonly excluded from the data set because can cause serious problems in statistical analyses. There are some methods and techniques to determine outliers such as Peirce's criterion or Thompson Tau and Modified Thompson Tau test. Other methods based on observations based are based on measures such as the interquartile range. For example, if Q_1 and Q_3 are the lower and upper quartiles, respectively, then we could define an outlier as an observation outside the range:

$$\left[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)\right], \tag{1}$$

for some nonnegative constant k. John Tukey proposed this test, where $k = 1.5$ indicates an "outlier", and $k = 3$ indicates data that is "far out".

Currently, outlier detection using the previous formula is used to determine if the mini-term is not working properly. The procedure to use is as follows:

- When a mini-term is registered, an initial calibration is carried out using an initial set of mini-terms $\{m_1, \ldots, m_t\}$ to determine how the machine or component performs in normal operation by adjusting the interquartile range. During the registration process, the operator has a chart of the data to decide if it follows a normal distribution. If two overlapping normal distributions appear in that chart, the mini-term would be considered as "programming error".
- The limits of the alarms are defined with the calculation of the quartiles using the initial set of mini-terms $\{m_1, \ldots, m_t\}$. If the anomaly is in the range $k = [1.5 - 3]$ it is defined as a *Warning* while if the anomaly is in the range > 3 the alarm is defined as type *Red*.
- If an alarm occurs and it is classified as *Warning*, an e-mail is sent to the head of maintenance, who decides if the variation is considered sufficient to be sent to the maintenance operators.
- If an alarm occurs and it is classified as *Red*, an e-mail is sent directly to the maintenance operators who will check the component.

Figure 9 shows some examples of mini-terms where 1 is an alarm that happened in a clamp, 2 is an alarm in a rollertable, 3 is in a lifter and 4 in a welding gun.

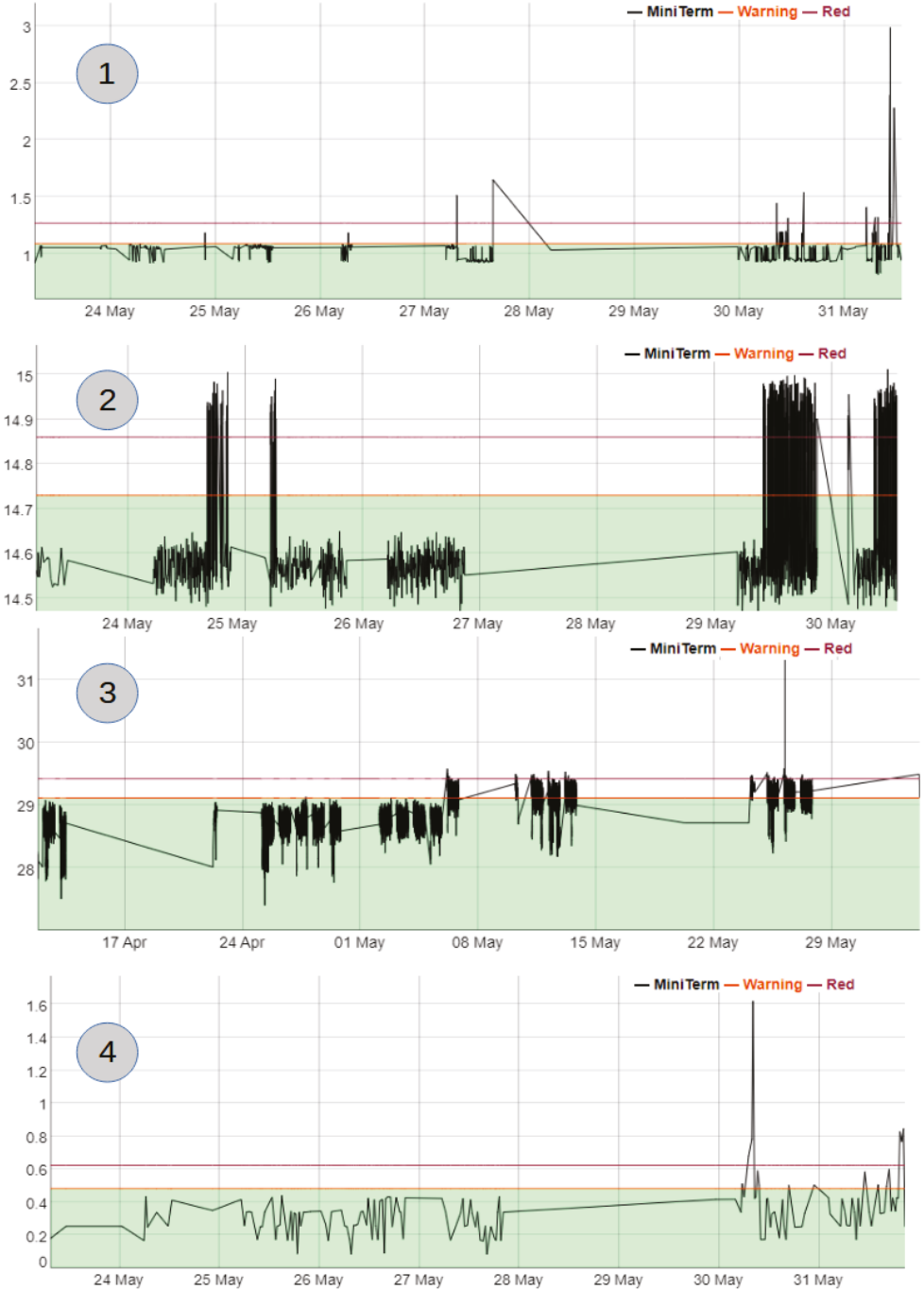


Figure 9. Examples of mini-term alarm. (**1**) Clamp, (**2**) Rollertable, (**3**) Lifter, (**4**) Welding gun.

6. Experimental Results

The process of installation and monitoring of machines and components began in 2018, in which 200 mini-terms were installed. In 2019, we were granted the project funded by the CDIT (IDI-20190878) (Centre for the Development of Industrial Technology), a Public Business Entity, under the Ministry of Science and Innovation which promotes innovation and technological development for Spanish companies. This project made it possible to massively implement the mini-terms at the Ford Factory in Almussafes (Valencia) and develop the necessary algorithms for the detection of machine failures. To date, June 2022, there are 18,893 mini-terms installed, of which 15,009 are pneumatic cylinders, 427 are pneumatic welding clamps, 24 pumps, 131 pantographs, 216 fans, 123 anchor matrices, 597 servos, 1690 tables and 479 rotary tables, 161 switches and 36 reducers.

6.1. Effectiveness of the Detection Algorithm

As explained in the previous section, the fault detection algorithm has undergone different evolutions. Initially, an algorithm based on K-Means was developed [21], which made it possible to collect cases of faults and carry out a more in-depth study of change point detection algorithms, see [24], where the conclusion was that Bartlett's algorithm was the most suitable one. However, this algorithm still had problems when the anomaly was fluctuating, or slow, and also detected the Scan-Time as an anomaly. To solve it, in this article an algorithm based on outlier detection from an initial calibration has been proposed. Although the proposed algorithm is very simple, it is extremely robust because in the initial calibration process, it is able to absorb the Scan-Time of the mini-term and it will not be detected as an anomaly. In addition, the slow deterioration can be detected as it starts from an initial calibration which can find any slow deterioration. The same is applied to oscillation pathologies. Thus, the effectiveness limit mentioned in [7] has been reached, where it was indicated that 99% of equipment failures are preceded by certain signs, conditions or indications that the failure is about to occur. That 1% remaining corresponds to components that have a sudden failure.

6.2. Benefits of Using Mini-Terms in Industry

Since the mini-terms began to be installed and used to predict breakdowns, different types of indicators have begun to undergo significant improvements, including:

- Technical availability (TAV): Percentage of planned production time without unexpected technical difficulties or maintenance needs.
- Mean Time To Repair (MTTR): Average time required to repair a failed component or device.
- Mean Time Between Failure (MTBF): Elapsed time between inherent failures of a mechanical or electronic system, during normal system operation.
- Number of Work order EM (Emergency Orders/line Stop).
- Labor hours in EM (Emergency Orders/line Stop).

Figure 10 shows the evolution of these indicators between the years 2016–2021 and compares it with the number of mini-terms installed. Due to industrial data protection issues, these graphs do not show the actual values of each indicator but instead they show the incremental improvement of each indicator since the initial year, 2016. In this comparison, and the effects that mini-terms cause, we need to take into account the improvement of the algorithm for detecting the change points and its increased reliability over the years. With respect to TAV, it has undergone an improvement of 18% from 2016 to 2021. In the case of the MTTR, it has been reduced by 2 h. Regarding the number of work orders, these have been significantly reduced thanks to the prediction of failures, eliminating about 3000 orders/year, while working hours have also been drastically reduced by almost 4000 h/year.

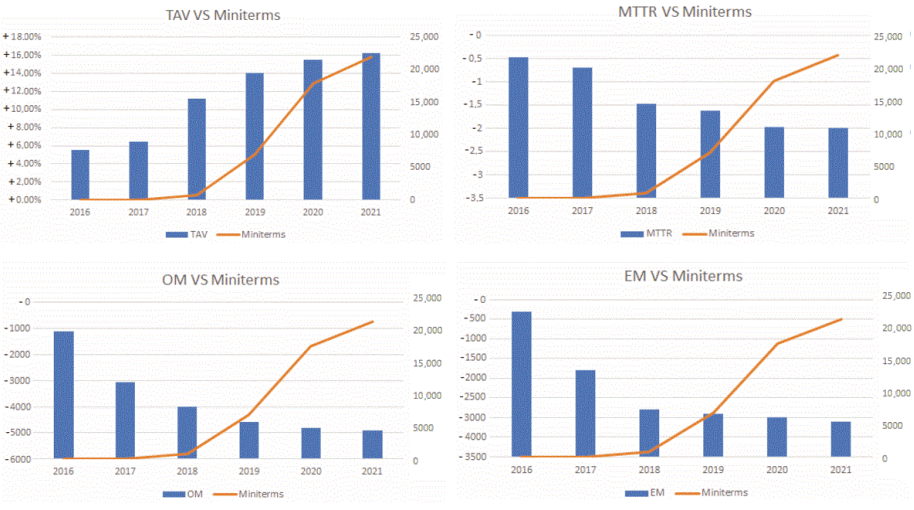


Figure 10. Evidence of the benefits of using mini-terms in production. NOTE: Due to industrial data protection issues, these graphs do not show the actual values of each indicator but instead they show the incremental improvement of each indicator since the initial year, 2016.

7. Discussion

The appearance of mini-terms as virtual sensors used for predictive maintenance has become a huge development for the analysis of the behaviour of components and machines. Their main advantage is that their installation does not involve a large financial outlay and therefore Big Data can be developed for a wide variety of components and machines. Currently, as predictive maintenance requires the installation of sensors, etc., the industry often chooses critical components or machines for its installation. However, a line stop can be caused by a component as small as a switch. The mini-term allows to monitor the health of this type of components in a global way in an entire factory.

There are several limitations when using mini-terms in the industry. These include:

- The use of the industrial network for data transmission and PLCs to measure cycle times may cause certain technical limitations. These are:
  - When the number of mini-terms increases significantly, the industrial network may suffer and directly affect production, due to network saturation.
  - When the mini-term value, cycle time to be measured, is small and approaches or exceeds the PLC cycle time (Scan-Time), it generates distortions in the data and the change point can be masked within the effect generated by the Scan-Time.
- The use of the mini-terms is based on the variation of the cycle time due to deterioration. However, when the element or component has a control system, it may hide this temporary deterioration. We can take as example a welding clamp with servomotor that begins to have a mechanical deterioration. These systems have a closing speed setpoint that the control system will try to keep at all costs, hiding the deterioration from the point of view of cycle time.

8. Conclusions and Future Work

This article introduces a novel virtual sensor for predictive maintenance called mini-term. The advantage introduced by the mini-term is that it can be implemented in an easy and simple way in pre-installed systems since you only need to program a timer in the PLC or PC that controls the line/machine in the production line, allowing, according to the authors' knowledge, to build industrial Big Data on predictive maintenance for the first time, which is called Miniterm 4.0. This article shows the current state of

development of the Miniterm 4.0 intelligent system where, the first objective is to build a global monitoring system for technical sub-cycle times (mini-terms) that allows predicting machine/component failures in real time. After installing more than 14,000 mini-terms in the last 3 years at Ford factory in Almussafes (Valencia), evidence has been generated of how the Miniterm 4.0 system can improve production rates. Indicators such as TAV (Technical availability), Mean Time To Repair (MTTR), Mean Time Between Failure (MTBF), EM (Number of Work order (Emergency Orders/line Stop)) and OM (Labour hours in EM) have generated a very important improvement as the number of mini-terms increased and the Miniterm 4.0 system became more reliable. In particular, TAV is increased by 15%, OM is reduced in 5000 orders, MTTR is reduced in 2 h and there are produced 3000 orders less than when mini-terms did not exist.

As future work, we will intend to complement the function of the mini-term as a predictor of failures functioning as a sub-cycle time, allowing to estimate the real capacity of production lines, as well as estimating the loss due to deterioration, determining the bottleneck in dynamic, etc. A first approach to the benefits that this new capacity can bring to the Miniterm 4.0 system has been presented in [29].

Author Contributions: Formal analysis, N.M.; Funding acquisition, E.G. and N.M.; Investigation, E.G., N.M., J.L. and A.L.; Methodology, E.G., N.M., J.L. and A.L.; Project administration, E.G.; Resources, N.M.; Software, J.L. and A.L.; Supervision, J.L. and A.L.; Validation, E.G., J.L. and A.L.; Visualization, J.L. and A.L.; Writing—original draft, N.M.; Writing—review and editing, E.G. and N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CDIT(Centre for the Development of Industrial Technology) with number IDI-20190878.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Battaïa, O.; Dolgui, A. A taxonomy of line balancing problems and their solution approaches. *Int. J. Prod. Econ.* **2013**, *142*, 259–277. [CrossRef]
2. Ben-Daya, M.; Duffuaa, S.O. Maintenance and quality: The missing link. *J. Qual. Maint. Eng.* **1995**, *1*, 20–26. [CrossRef]
3. Montgomery, D.C. *Introduction to Statistical Control*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1985.
4. Khan, M.R.; Darrab, I.A. Development of an analytical relation between maintenance, quality and productivity. *J. Qual. Maint. Eng.* **2010**, *16*, 341–355. [CrossRef]
5. Falkenauer, E. Line Balancing in the Real World. In *International Conference on Product Lifecycle Management*; Interscience Enterprises Ltd.: Osaka, Japan, 2005.
6. Peng, Y.; Dong, M.; Zuo, M.J. Current status of machine prognostics in condition-based maintenance: A review. *Int. J. Adv. Manuf. Technol.* **2010**, *50*, 297–313. [CrossRef]
7. Bosch, H.P.; Geitner, F.K. *Machine Failure Analysis and Troubleshooting*; Gulf Publishing Co. Technology & Engineering: Houston, TX, USA, 1983; Volume 50.
8. Ahmad, R.; Kamaruddin, S. An overview of time-based and condition-based maintenance in industrial application. *Comput. Ind. Eng.* **2012**, *63*, 135–149. [CrossRef]
9. Jardine, D.A.K.S.; Banjevic, D.; A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signals Process.* **2006**, *20*, 1483–1510. [CrossRef]
10. Kumar, S.; Goyal, D.; Dang, R.K.; Dhami, S.S.; Pabla, B.S. Condition based maintenance of bearings and gears for fault detection. A review. *Mater. Today Proc.* **2018**, *5*, 6128–6137. [CrossRef]
11. Jeong, I.J.; Leon, V.J.; Villalobos, J.R. Integrated decision-support system for diagnosis, maintenance planning and scheduling of manufacturing systems. *Int. J. Prod. Res.* **2007**, *45*, 2007. [CrossRef]
12. Rastegari, A. *Condition Based Maintenance in the Maintenance Industry. From Strategy to Implementation*; Malardalen University: Västerås, Sweden, 2017.
13. Corcoran, J.; Davies, C.M. Monitoring power-law creep using failure forecast method. *Int. J. Mech. Sci.* **2018**, *140*, 179–188. [CrossRef]

14. Zhao, X.; Cai, K.; Wang, X.; Song, Y. Optimal replacement policies for a shock model with a change point. *Comput. Ind. Eng.* **2018**, *118*, 383–393. [CrossRef]
15. Nigro, M.B.; Packzad, S.N.; Dorvash, S. Localized structural damage detection. A change point. *Comput.-Aided Civ. Infrastruct. Eng.* **2014**, *29*, 416–432. [CrossRef]
16. Al-Kandari, N.M.; Aly, E.E.A.A. An ANOVA-type test for multiple change points. *Stat. Pap.* **2014**, *55*, 1159–1178. [CrossRef]
17. Killick, R.; Eckley, I.A. Changepoint: An R Package for Changepoint Analysis. *J. Stat. Softw.* **2014**, *58*, 1–19. [CrossRef]
18. García, E. Análisis de los sub-tiempos de ciclo técnico para la mejora del rendimiento de las líneas de fabricación. Ph.D. Thesis, Universidad CEU-Cardenal Herrera, Alfara del Patriarca, Spain, 2016
19. Li, L.; Chang, Q.; Ni, J. Real time production improvement through bottleneck control. *Int. J. Prod. Res.* **2009**, *47*, 6145–6158. [CrossRef]
20. Garcia, E.; Montés, N. A tensor model for automated production lines based on probabilistic sub-cycle times. In *Modeling Human Behaviour: Individials and Organizations*; Nova Science Pulishers: Hauppauge, NY, USA, 2017; pp. 221–234.
21. Garcia, E.; Montés, N.; Alacreu, M. Towards a Novel Maintenance Support System Based On mini-terms: Mini-term 4.0. In *Informatics in Control, Automation and Robotics, Proceedings of the ICINCO, Porto, Portugal, 29–31 July 2018*; Gusikhin, O., Madani, K., Eds.; Lecture Notes in Electrical Engineering; Springer: Cham, Switzerland, 2019; Volume 613, pp. 101–117.
22. Llopis, J.; Lacasa, A.; Garcia, E.; Montés, N.; Hilario, L.; Vizcaino, J.; Vilar, C.; Vilar, J.; Sánchez, L.; Latorre, J.C. Manufacturing Maps, a Novel Tool for Smart Factory Management Based on Petri Nets and Big Data Mini-Terms. *Mathematics* **2022**, *10*, 2398. [CrossRef]
23. Killick, R. Package 'Changepoint'. 2016. Available online: https://cran.r-project.org/web/packages/changepoint/changepoint.pdf (accessed on 14 August 2022).
24. Garcia, E.; Montés, N.; Llopis, J.; Lacasa, A. Evaluation of Change Point Detection Algorithms for Application in Big Data Mini-term 4.0. In Proceedings of the International Conference on Informatics in Control, Automation and Robotics, ICINCO, Paris, France, 7–9 July 2020.
25. Bect, P.; Simeu-Abazi, Z.; Maisonneuve, P. Identification of abnormal events by data monitoring: Application to complex systems. *Appl. Complex Syst. Comput. Ind.* **2015**, *68*, 78–88. [CrossRef]
26. Lei, Y.; Zuo, M.J. Gear crack level identification based on weighted Knearest neighbor classification algorithm. *J. Mech. Syst. Signal Process.* **2009**, *23*, 1535–1547. [CrossRef]
27. Tsui, K.L.; Chen, N.; Zhou, Q.; Hai, Y.; Wang, W. Prognostics and Health Management: A Review on Data Driven Approaches. *Math. Probl. Eng.* **2015**, *2015*, 793161. [CrossRef]
28. Wang, D. K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited. *Mech. Syst. Signal Process.* **2016**, *70*, 201–208 [CrossRef]
29. Garcia, E.; Montés, N.; Rosillo, N.; Llopis, J.; Lacasa, A. A novel model to analyse the effect of deterioration on machine parts in the line throughput. In Proceedings of the International Conference on Informatics in Control, Automation and Robotics, ICINCO, Paris, France, 7–9 July 2020.

sensors

MDPI

Article

A Novel Method of Impeller Blade Monitoring Using Shaft Vibration Signal Processing

Jindrich Liska, Vojtech Vasicek * and Jan Jakl

NTIS—New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitni 8, 301 00 Plzen, Czech Republic; jinliska@ntis.zcu.cz (J.L.); jjakl@ntis.zcu.cz (J.J.)
* Correspondence: vasicekv@ntis.zcu.cz

Abstract: The monitoring of impeller blade vibrations is an important task in the diagnosis of turbomachinery, especially in terms of steam turbines. Early detection of potential faults is the key to avoid the risk of turbine unexpected outages and to minimize profit loss. One of the ways to achieve this is long-term monitoring. However, existing monitoring systems for impeller blade long-term monitoring are quite expensive and also require special sensors to be installed. It is even common that the impeller blades are not monitored at all. In recent years, the authors of this paper developed a new method of impeller blade monitoring that is based on relative shaft vibration signal measurement and analysis. In this case, sensors that are already standardly installed in the bearing pedestal are used. This is a significant change in the accessibility of blade monitoring for a steam turbine operator in terms of expenditures. This article describes the developed algorithm for the relative shaft vibration signal analysis that is designed to run in a long-term perspective as a part of a remote monitoring system to track the natural blade frequency and its amplitude automatically.

Keywords: steam turbine; impeller blade; vibration; monitoring; diagnostics; algorithm; signal processing

Citation: Liska, J.; Vasicek, V.; Jakl, J. A Novel Method of Impeller Blade Monitoring Using Shaft Vibration Signal Processing. *Sensors* **2022**, *22*, 4932. https://doi.org/10.3390/s22134932

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 13 May 2022
Accepted: 28 June 2022
Published: 29 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The growing demand for higher power generation brings the need for increasing reliability and also the efficiency of steam and gas turbines together with ensuring safe operation. Early detection of potential fault is the key to avoid the risk of unexpected turbine outages and to minimize profit loss. One of the critical parts of the turbine is the blades of the last stages of the low-pressure turbine, where acting forces are essential in terms of the residual life. This is the key part that needs to be monitored in a long-term perspective to ensure the warning if any potential fault is present. In general, the existing approaches to blade vibration monitoring are based on contact or non-contact measurement.

The contact approach is based on strain gauging. This provides information about the mechanical stress on the blade surface. Because of the contact principle, it allows to obtain very accurate stress sampled at a high rate. This is useful in the signal analysis that follows. However, it is also necessary to transmit the electric signal outside the rotating part of the turbine [1]. In addition, the sensors must be resistant to the extreme conditions that occur in the flow section of the turbine. These are the reasons the strain gauge is unsuitable for the long-term monitoring of impeller blades.

The non-contact method can be represented by the popular blade tip-timing method (BTT) [2]. Sensors, in that case, are built right into the stator body to detect the times of the arrival of the blades. Detected times of arrival are further processed to evaluate the blade tip deflection. Vibrations of each blade are sampled once per revolution and, thus, proper diagnostics and condition monitoring require specialized signal processing techniques to be applied. In principle, this method is suitable for long-term monitoring. However, the costs of specialized sensors and the need for their installation right into the stator body are

the main reasons that BTT systems are not so widespread. Additionally, the installation costs are not negligible. It is even a frequent practice not to monitor the impeller blades at all.

Another technique for measuring blade vibrations mentioned in the literature is the use of a pressure sensor installed in the stator body [3,4]. In principle, the rotating blades cause pressure fluctuations in the inner part of the stator, especially between the tip of the blade and the wall of the stator. The frequency of this fluctuation is equal to the blade passing frequency (BPF). The location of installed sensors can be compared to the BTT method. Unlike BTT, the measurement is, in principle, not limited by the sampling frequency given by the shaft speed; however, the need for the installation of the specialized sensors right into the stator body remains. The impeller, as mentioned in the previous paragraph, causes pressure fluctuations on the inside of the stator by its rotation. The pressure fluctuation causes a force to act on the stator wall and, thus, the absolute stator vibrations can be used for blade vibration monitoring. The advantage of this approach is, in contrast to the pressure measurement inside the stator, the relative simplicity of the sensor installation. However, this approach does not provide the direct measurement of blade vibrations. The location of the accelerometer is usually above the impeller, or the standard sensors installed on bearing pedestals are used. In that case, the monitored BPF amplitude may change by propagating the vibrations through the material.

Recently, the use of torsional rotor vibrations in terms of blade vibration monitoring was published [5]. Torsional rotor vibrations can be characterized as angular oscillations of the shaft. In principle, the reference markers around the shaft are measured in each revolution. The reference can be the zebra tape or, for example, the teeth of the gear. The measured signal has a character of a pulse signal similar to times of arrival in the case of BTT. It turns out that the frequency spectrum of such a signal also contains information about blade vibrations. The physical principle of how the blade oscillations propagate to the measured signal is described in [6]. It states that the total moment of inertia given by the contribution from all blades must be large enough to be reflected on the shaft. It is obvious that the contributions of the individual blades cancel each other out in the total moment of inertia. The exception for the case of tangential natural vibrations of the blades, which can cause torsional vibrations of the shaft, is 0 nodal diameter, when all blades oscillate with the same phase. Until recently, only 0 nodal diameters of blade vibrations were captured using the torsional vibration measurement.

2. Method of Impeller Blade Monitoring

A novel approach to impeller blade monitoring was developed in recent years by the authors of this paper. This approach is based on the evaluation of the blade vibrations using the shaft vibration signal analysis. The measurement of shaft vibrations is made using the standardly installed sensors placed in the bearing pedestals. This is a significant change in the accessibility of impeller blade monitoring for a steam turbine operator in terms of expenditures. There is no need to install any special sensor. Blade vibrations are evaluated using the already installed standard instrumentation and there is no need for the steam turbine outage. The sensor location for this approach together with the comparison with BTT sensors placing is illustrated using the low-pressure turbine scheme in Figure 1.

The blade vibration is present in shaft vibrations because of the bending moment acting on the shaft that is caused by blade axial oscillations. In fact, the blade vibrations are present in a form of two spectral components. The frequencies of these components are described by $f_{rot} \pm f_n$. This is due to an amplitude modulation with a suppressed carrier that occurs under the sensor during the rotation. These two components are the so-called lower sideband (LSB) and upper sideband (USB), as shown in Equations (1) and (2) respectively:

$$\mathrm{USB} = f_{rot} + f_n \tag{1}$$

$$\mathrm{LSB} = f_{rot} - f_n \tag{2}$$

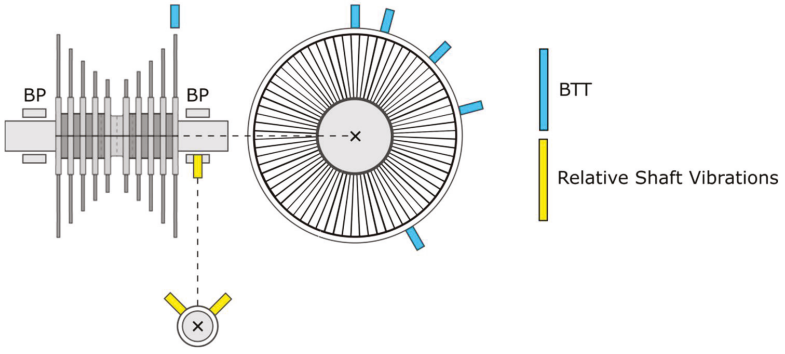


Figure 1. Low-pressure turbine scheme with the sensor location of BTT sensors and standard shaft vibration sensors.

The illustrations of blade components are illustrated in Figure 2, where on the left is the spectrogram of shaft vibration signal measured in real operation of the low-pressure turbine. On the right is the same spectrogram with highlighted LSB and USB components related to the blade vibration for the first three bending modes (BM). From this Figure, the presence of blade vibration components is clear.

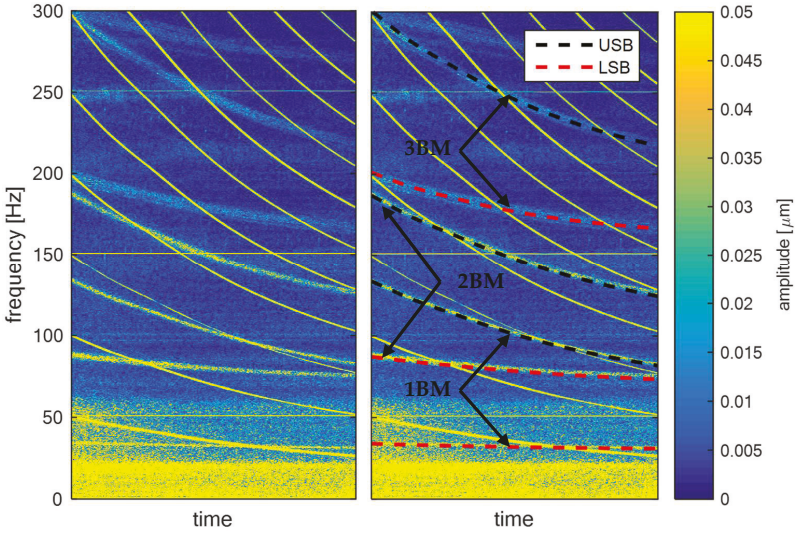


Figure 2. Spectrogram of shaft vibration signal from low-pressure turbine (**left**) with highlighted LSB and USB components (**right**).

The detailed description of this principle was published in the literature recently [7–10]. For proper impeller blade state diagnostics, the exact parameters of the blade components from the shaft vibration signal need to be obtained. Such a method is not published yet, so the purpose of this research was to develop and introduce the appropriate method for real-time blade monitoring using the shaft vibration signal from the long-term perspective. The description of this method is the main objective of this article.

Firstly, the measured signal is pre-processed. This includes the signal filtering based on cepstral liftering and the spectrum time averaging in the spectral plane (flowchart in Figure 3) where the block schema of the proposed method is illustrated. The pre-processed signal is then analyzed and the blade components are identified, filtered in the time-

frequency domain and clustered using the Euclidean distance. This algorithm is designed to run in a long-term perspective as a part of a remote monitoring system to track the natural blade frequency and corresponding amplitude automatically.

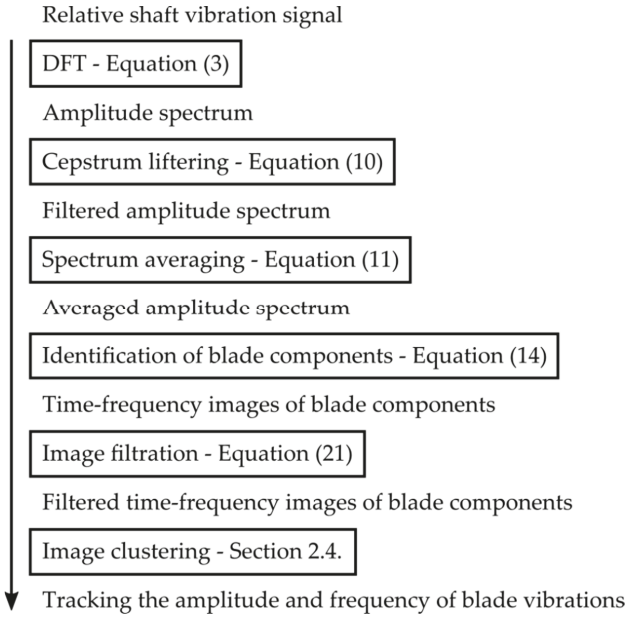


Figure 3. Flowchart of the developed method for online long-term blade vibration monitoring.

2.1. Cepstral Liftering of the Shaft Vibrations

The measured time signal of relative shaft vibrations is pre-processed first. The pre-processing is made in the frequency domain. For this purpose, the Discrete Fourier transform (DFT) is used because the measured signal is finite and discrete [11–14]. Formula (3) defines a spectrum of a signal $x[n]$ sampled by frequency f_s, where n is the index of the signal sample:

$$X[k] = \mathrm{DFT}\{x[n]\} \tag{3}$$

At this point, it is also desirable to define a one-sided amplitude spectrum that will be used afterwards (Formula (4)). The spectral frequency corresponding to integer index k follows Formula (5), where L is the length of the signal.

$$A[k] = \begin{matrix} |X[k]|, & k = 0 \\ 2|X[k]|, & k > 0 \end{matrix} \tag{4}$$

$$f_k = k \cdot f_s / L \tag{5}$$

Measurement of the relative shaft vibration signal contains the noise that is unevenly distributed across the amplitude spectrum. To be able to identify the blade vibrations components using the following analysis, the signal noise needs to be filtered first. For this purpose, the cepstral analysis and spectral envelope calculation are used [15].

The cepstrum calculation is the basic method of cepstral analysis. The so-called real cepstrum that is defined by Formula (6) is an inverse Fourier transform (IDFT) of a logarithm of an amplitude spectrum of the sampled signal Formula (6).

$$c[n] = \mathrm{IDFT}\{\ln(|X[k]|)\} \tag{6}$$

The cepstrum itself characterizes the periodicity of the amplitude spectrum. Low quefrencies represent the low-frequency periodicity in the spectrum and vice versa. This can be used for cepstrum liftering. It is a process of weighting the cepstrum by the lifter function. This is represented by Formula (7):

$$c_w[n] = c[n] \cdot w[n] \tag{7}$$

Using liftering the unwanted cepstrum coefficients are suppressed. For the spectral envelope calculation, the lifter in a form of the Gauss function is used (Formula (8)). In that case, the high quefrencies are suppressed leaving just low quefrencies in cepstrum that can be used for the spectral envelope calculation (Formula (9)):

$$w[n] = e^{-\frac{n^2}{2\sigma^2}} \tag{8}$$

$$E[k] = e^{\mathrm{DFT}\{c_w[n]\}} \tag{9}$$

The level of the noise that is present in the measurement is then filtered simply by subtraction of the spectral envelope from the original amplitude spectrum (Formula (10)):

$$X_F[k] = A[k] - 2E[k] \tag{10}$$

2.2. Time Averaging of the Amplitude Spectrum

Filtering of the amplitude spectrum of the shaft vibration signal using the cepstrum calculation suppresses the mean value of the measurement noise on each spectral component. In addition to the mean value, the noise is also characterized by the variance in time that also needs to be suppressed. For this purpose, spectrum time averaging is used and so the frequency domain is extended into the time-frequency domain. The measured signal is divided into equidistant time series that are then transformed into a spectrum by DFT. The average amplitude spectrum is then defined using Formula (11) whereas N represents the number of averaged DFT spectrums. The time index l is tied up with the sampling frequency and parameter of the signal window shift in samples Δ_w from one DFT calculation to another (Formula (12)). For further use, the average amplitude spectrum is also indexed in time by index m similar to l. t_m represents the time of average amplitude spectrum that is based on the shift in samples Δ_m in a similar manner to index l (Formula (13)):

$$X_F^T[m,k] = \frac{1}{N}\left|\sum_{l=m}^{m+N-1} X_F[l,k]\right| \tag{11}$$

$$t_l = \frac{l \cdot \Delta_w}{f_s} \tag{12}$$

$$t_m = \frac{m \cdot \Delta_m}{f_s} \tag{13}$$

2.3. Identification and Filtration of the Blade Components

The pre-processed relative shaft vibration signal as a form of averaged cepstrum liftered amplitude spectrum is the input to the process of automatic identification of blade vibrations. The robustness of the calculation is taken into account because this algorithm is supposed to run on a long-term base with no user intervention. In this chapter, the process of the blade component identification together with its filtration is described. The expert task is also the proper selection of frequency interval to be monitored, defined by f_{MIN} and f_{MAX}. The assumption for obtaining correct results for the identification process is that the monitored interval must not include the other excited components that are not related to the blade vibrations on such higher harmonics, etc.

The spectral components that are defined as blade vibration components in a shaft vibration signal spectra follow Formula (14). It is a set of $X_F^T[t_m, f_k]$ that meets the conditions,

its frequency is in the range defined by f_{MIN} and f_{MAX} (see C_1 (15)) and its amplitude is higher than the multiple of parameter A_{TH}, defined by the expert and the median of $X_F^T[t_m, f_k]$ values around the analysed component, (see C_2 (16) and (17)). The size of the frequency interval for the median calculation represents parameter δ [Hz]:

$$X_{NF}[t,f] = \left\{ X_F^T[t_m, f_k] \middle| C_1 \wedge C_2 \right\} \tag{14}$$

$$C_1 : \; f_k \in \langle f_{MIN}, f_{MAX} \rangle \tag{15}$$

$$C_2 : \; X_F^T[t_m, f_k] > T[t_m, f_k] \tag{16}$$

$$T[t_m, f_k] = A_{TH} \cdot \text{med}\left(X_F^T[t_m, f] \middle| f \in \langle f_k - 0.5 \cdot \delta, f_k + 0.5 \cdot \delta \rangle \right) \tag{17}$$

The components identified according to (14) are then filtered to suppress the events in a short-term horizon that do not follow the blade behavior. For this purpose, a subset of (14) for every component from (14) is defined according to Formula (18). This subset contains all $X_{NF}[t,f]$ from the rectangle $R[t,f]$ defined in the time-frequency domain according to conditions C_3 and C_4 (see Formulas (19) and (20), respectively):

$$R[t,f] = \{ X_{NF}[t,f] | C_3 \wedge C_4 \} \tag{18}$$

$$C_3 : \; t \in \langle t, t + t_{FILT} \rangle \tag{19}$$

$$C_4 : \; f \in \langle f - 0.5 \cdot f_{FILT}, f + 0.5 \cdot f_{FILT} \rangle \tag{20}$$

The example of this rectangle is illustrated for X_{NF} [57.5, 70.6] and X_{NF} [57.5, 72] in red in Figure 4. The meaning of the parameters t_{FILT} and f_{FILT} from this figure is evident—it is the time length and frequency length, respectively, of window rectangle $R[t,f]$. The red dots represent other identified $X_{NF}[t,f]$ according to the algorithm described in a previous part of this article. The filtration itself is defined according to Formula (21). The non-filtered $X_{NF}[t,f]$ is filtered if the number of $X_{NF}[t,f]$ in $R[t,f]$ does not exceed the threshold N_{TH}. From Figure 4, it is obvious that $X_{NF}[57.5, 70.6]$ becomes $X_I[57.5, 70.6]$ and $X_{NF}[57.5, 72]$ not because $|R[t,f]| = 0$. This process filters randomly excited noise components that can randomly occur during the measurement in the short-term, making this algorithm more robust and to ensure that false alarms are not generated.

$$X_I[t,f] = \{ X_{NF}[t,f] | N_{TH} > |R[t,f]| \} \tag{21}$$

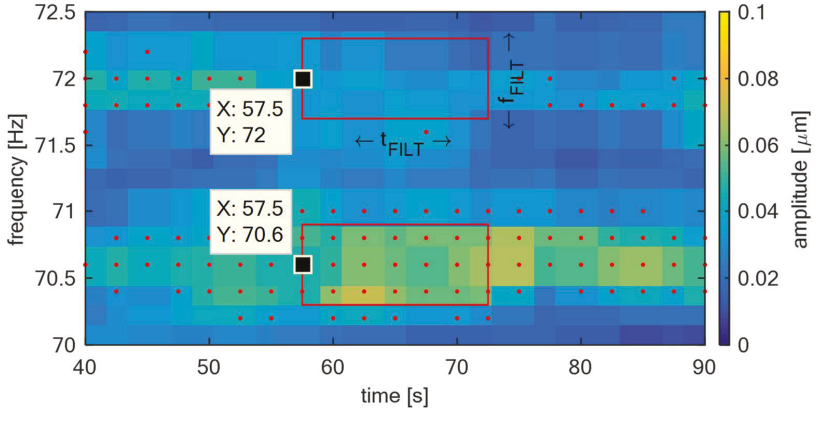


Figure 4. TG 250 MW—Filtering of identified components in the spectrum.

2.4. Clustering of the Blade Components

For practical use, the identified blade vibration components from the previous chapter need to be interpreted in an intelligible way. It is appropriate to represent each of the blade frequencies with one frequency and amplitude value that can be tracked over time, instead of many identified components. For this purpose, the identified blade components are clustered using cluster analysis [16].

This analysis groups identified components into sets according to a defined degree of similarity. The red components identified in the previous section and illustrated in Figure 4 are being called images of blade vibrations in the time-frequency plane of the relative shaft vibration signal. Using this terminology, the i-th image is indexed and abbreviated according to Formula (22). The degree of similarity of two images is based on their frequency distance. This can be defined by the Euclidean distance represented by operator d that is in form of the Formula (23) for one-dimensional space. This relation defines the distance between the i-th and j-th image X_I:

$$X_I^i = X_I[t_i, f_i] \tag{22}$$

$$\mathrm{d}\left(X_I^i, X_I^j\right) = |f_i - f_j| \tag{23}$$

The clustering process used in this algorithm is defined according to the following sequence. In the first step of clustering at time t_1, image 1 is declared to be the first cluster (Formula (24)). The cluster is given by S and indexed by k. Each cluster can be represented by its center. The center of cluster S_k is defined using Formula (25), where N_k denotes the number of images X_I assigned to the cluster S_k:

$$X_I^1 \in S_1 \tag{24}$$

$$\mu_k = \frac{1}{N_k} \sum_{f \in S_k} f \tag{25}$$

In the following j-th iteration of this algorithm, steps represented by Formulas (26)–(30) are being repeated for the analyzed X_I^j image. The first procedure is to exclude the images from the cluster sets that are older than the time of the actual image for more than the forgetting factor t_{TH} $[s]$ (see Equation (26)). This step is important to allow the algorithm to capture any changes in blade natural frequency that occurs over the monitored period:

$$S_k = S_k \backslash \left\{ X_I |\, t < \left(t_j - t_{TH}\right) \right\} \tag{26}$$

Updating the cluster images is followed by the recalculation of their centers. The distance between the cluster centers and the j-th image X_I^j is then determined. The number of the clusters is K. The nearest cluster to the j-th image is I_{MIN} in a distance M_{MIN} (see Formulas (27) and (28)). If the distance M_{MIN} is less than the frequency threshold f_{TH}, then the j-th image is assigned to the I_{MIN}-th cluster (Formula (29)). Otherwise, a new cluster is established (Formula (30)) and the next iteration of the clustering process is executed. The choice of f_{TH} allows to set the precision of spectral components that can be distinguished by the proposed algorithm:

$$I_{MIN} = \underset{k}{\operatorname{argmin}}\, \mathrm{d}\left(X_I^j, \mu_k\right),\ k = 1, \ldots, K \tag{27}$$

$$M_{MIN} = \min_k \mathrm{d}\left(X_I^j, \mu_k\right),\ k = 1, \ldots, K \tag{28}$$

$$X_I^j \in S_{I_{MIN}},\ M_{MIN} \leq f_{TH} \tag{29}$$

$$X_I^j \in S_{k+1},\ M_{MIN} > f_{TH} \tag{30}$$

3. Results

To demonstrate the described algorithm, the measurement made in the operation of TG 250 MW was used. The shaft vibration signal was captured by an eddy current displacement sensor that met the standard for use in practice. The measuring range of the sensor was from 0–2 mm and the sensitivity 8 V/mm. The data acquisition was made using the National Instruments hardware. The Chassis cDAQ-9189 equipped with an NI-9229 4 channel module was used. This module allowed to measure analog inputs with 24-bit resolution, the maximal sampling frequency of 50 kHz and an input voltage range from −60 V to 60 V. The shaft vibration signal was captured using the sensor installed in the front bearing pedestal of the low-pressure turbine. The last stage blades were shrouded on the tips and in the middle with a tie-boss. The distance between the bladed wheel and the sensor was 2.4 m.

The example of DFT of the shaft vibration signal that was measured in the operation of TG 250 MW is illustrated in the form of an amplitude spectrum according to Equation (4) in Figure 5 in blue. The length of the analyzed signal was 5 s, the sampling frequency was 10240 Hz, the rotational frequency was 3000 rpm (nominal speed) and the Hanning window was used. In the signal spectrum, 1X frequency—50 Hz—and its integer multiplies were dominant. As well as those components, the blade vibration components can also be, principally, present in the signal. In Figure 5, there are two parts of the amplitude spectrum of the shaft vibration signal. At this point of the signal processing procedure, it is not clear if any of those two intervals also contain the blade vibrations or not. The calculated spectral envelope is illustrated by the dashed black line in Figure 5. For its evaluation, the parameter σ was set to 0.16 (see Equation (6)).

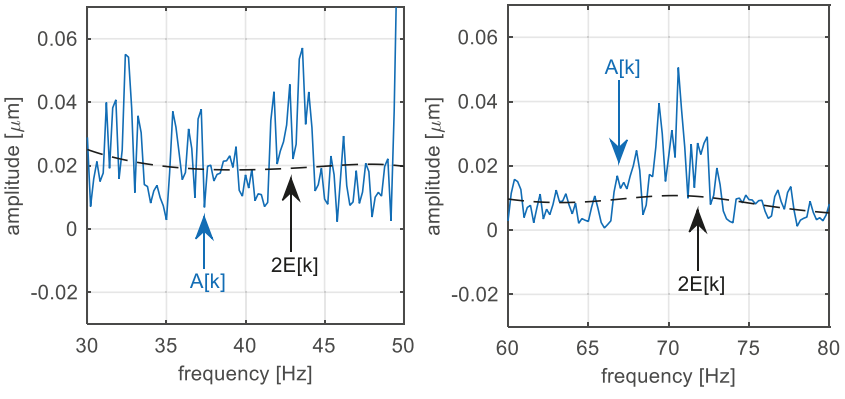


Figure 5. TG 250 MW—Discrete Fourier transform amplitude spectrum of the shaft vibration signal and its spectral envelope—liftered spectrum.

The step of the noise reduction according to Equation (10) is illustrated using Figure 6. It can be compared with the non-filtered amplitude spectrum in Figure 5. The noise level is suppressed by the spectral envelope. After this operation, the mean value of the spectral measurement noise on every spectral component is supposed to be suppressed.

Time averaging described in Section 2.2 was applied to the measurement, which is illustrated in Figure 7 in the frequency domain. The DFT is then defined in time instants corresponding to $\Delta_w = 5f_s$ (Equation (12)). The number of spectrums for averaging satisfies the condition $N = 20$, which means that the averaged spectrum characterizes 100 s of the relative shaft vibration signal.

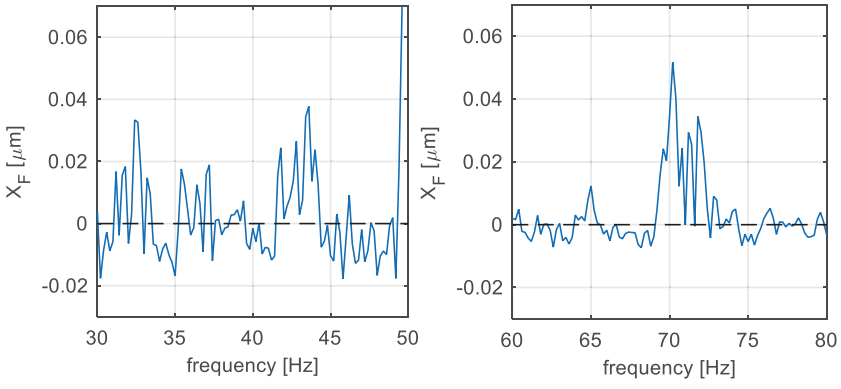


Figure 6. TG 250 MW—Filtered Discrete Fourier transform amplitude spectrum of the shaft vibration signal.

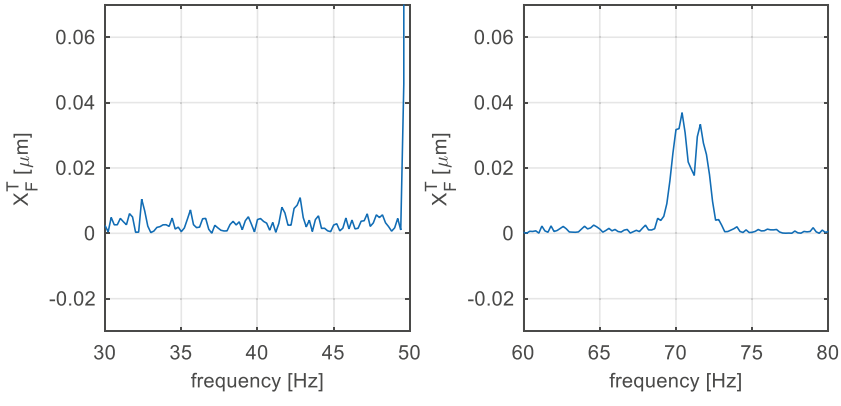


Figure 7. TG 250 MW—Time-averaged filtered amplitude spectrum of the relative shaft vibration signal.

In Figure 7, there are the same two amplitude spectrum intervals as in Figures 5 and 6. It can be seen from this Figure that the range from 58–83 Hz contains the excited components, while the range from 25–50 Hz contains only the noise. This was not obvious in the previous steps illustrated by Figures 5 and 6, respectively. The decision of whether the excitation is related to blade vibrations can be made by using a priori knowledge of the natural blade vibration frequencies standardly illustrated by the Campbell diagram [17]. Using this diagram, the approximate location of the blade natural frequency can be obtained and compared with the spectral components excited in a real measurement by the expert.

The example of the identification process is illustrated in Figure 8. There is the preprocessed signal of the shaft vibration in the form of an average amplitude spectrum, as it is illustrated in Figure 7. The dashed line represents the threshold defined in (17), that is in this case defined for $A_{TH} = 5$ and $\delta = 20$ Hz. The black squares illustrate the identified components that exceed the threshold satisfying condition (16). It can be seen in Figure 8, that, if only the noise is present in the signal spectra monitoring range, there are no identified blade components and vice versa. The parameters should be set uniquely for each monitored frequency interval.

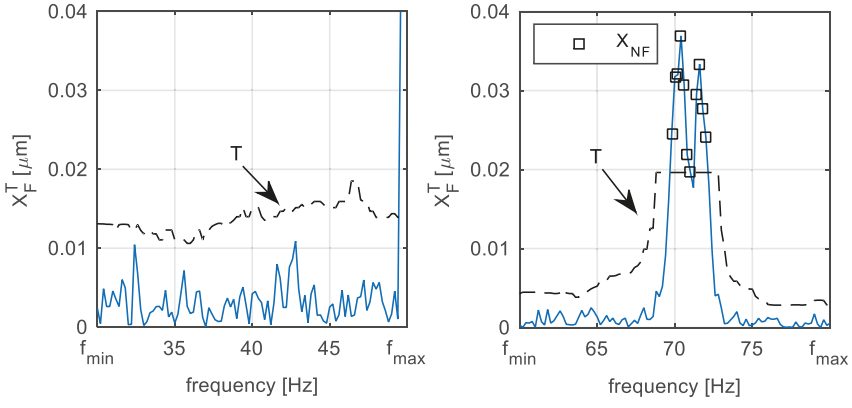


Figure 8. TG 250 MW—Blade component identification.

The example of the novel method for blade vibration monitoring that was described in this article is illustrated in Figures 9 and 10. In Figure 9, left, there is the spectrogram of shaft vibration signal measured in the operation of TG 250 MW. The sampling frequency of the measurement was 10.24 kHz, the length of the Hanning window used in DFT was 5 s and Δ_w was $2.5f_s$. It can be seen that the relative shaft vibration signal itself is noisy. In Figure 9, right, there is a spectrogram of the same signal section, but this time pre-processed according to Equation (11). It can be seen that the noise was filtered and the blade components that were approximately 71 Hz and 72.5 Hz are more obvious, even with the naked eye. The parameters for pre-processing were: $\sigma = 0.16$, $\Delta_w = \Delta_m$ and $N = 5$, which corresponds to 15 s of signal for averaging. In fact, there are two blade components that increase their amplitude in time and also change their frequency because of the change in the steam turbine operating conditions.

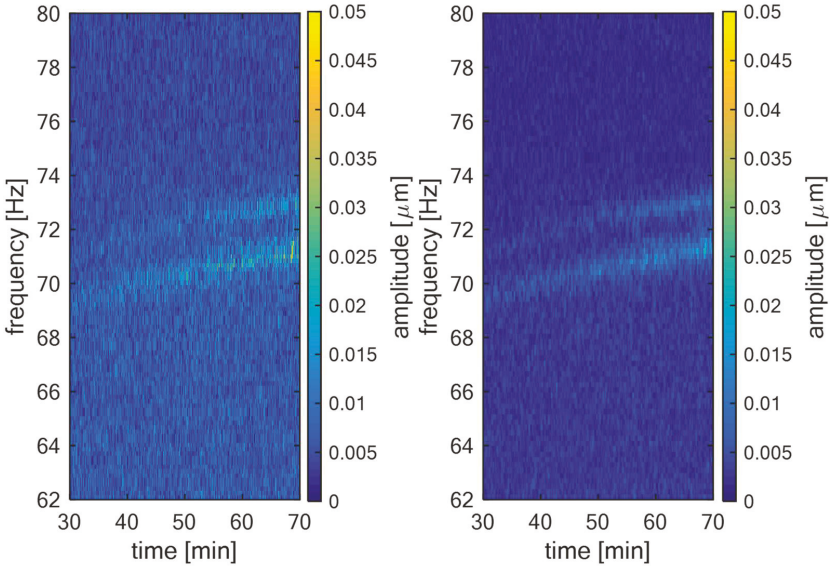


Figure 9. TG 250 MW—Amplitude spectrogram of the relative shaft vibration signal (**left**); averaged and filtered amplitude spectrogram of the same signal (**right**).

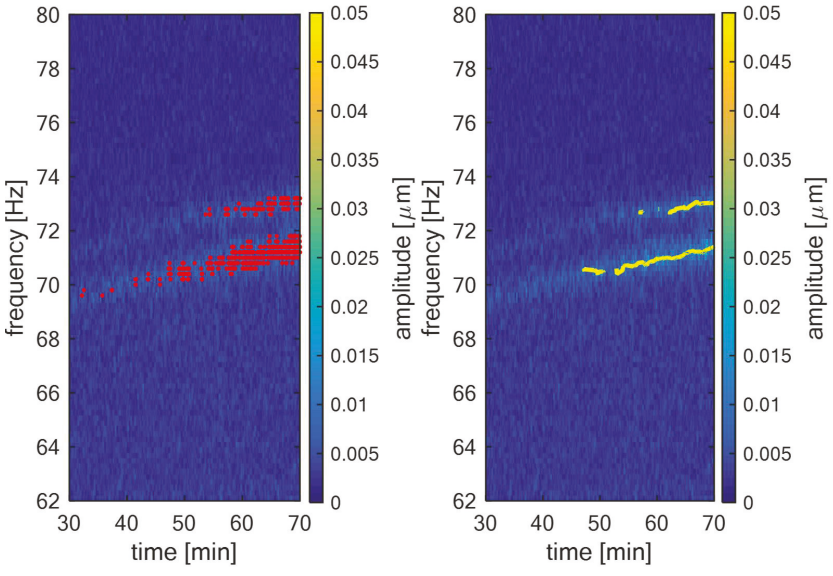


Figure 10. TG 250 MW—Identified blade vibration components in amplitude spectrogram of the relative shaft vibration signal (**left**); tracking of blade vibration components using the cluster analysis (**right**).

In Figure 10, left, the identified blade components according to Equation (14) are illustrated in red. The parameters for this step were: $\delta = 10$ Hz, $A_{TH} = 3$, $f_{MIN} = 62$ Hz and $f_{MAX} = 80$ Hz. In Figure 10, right, there is the trend of the blade components after the cluster analysis according to Equation (25), illustrated in yellow. The parameters for this step were: $t_{FILT} = 10$ s, $f_{FILT} = 0.8$ Hz and $N_{TH} = 8$. It can be seen that the developed algorithm effectively tracks both blade components that are present in the relative shaft vibration signal.

4. Discussion and Conclusions

The main objective of this work was to develop and validate the novel method for automatic long-term blade vibration monitoring using the relative shaft vibration signal analysis. This measurement uses the sensors that are standardly installed in the steam turbine. This is the main advantage over the other methods and principles of blade vibration monitoring. This approach is potentially interesting for steam turbine operation because there is no need of any special sensor installation. This article describes the developed method in detail. The measured vibration signal of shaft vibrations is firstly pre-processed. The spectral noise is filtered using the cepstrum analysis described in Section 2.1. Then, the spectral components are averaged in time to reduce the noise variance (see Section 2.2). After that, the pre-processed signal is used to identify the blade components that are excited according to the proposed identification rule described in detail in Section 2.3. The cluster analysis is the last step, which is used to merge blade components with close identified frequencies (see Section 2.4). Those frequencies may be tracked over time together with corresponding amplitudes and may be compared with the nominal values. If the differences reach the maximal allowed limit, then the alarm is triggered, meaning the blade state is in a non-standard state.

Another benefit of this approach is the easy acquisition of the natural frequencies of the installed blades from real operation without the need for outage. By simply connecting to existing standard instrumentation, the application can evaluate the actual natural frequencies of the blade, which differ slightly from numerical calculations. The turbine

operator can include this information to minimize the operation time for the rotational frequency, or its integer multiple is close to any of the resonant frequencies of the blades that was evaluated using proposed approach.

The algorithm itself was tested and validated on the measurement data from the turbo generator of 250 MW. The validation of this algorithm is illustrated in the figures and spectrograms in this article. After the validation phase, the algorithm was implemented into the online monitoring and diagnostics system available for commercial purposes, which is currently monitoring the operation of the 215 MW turbo generator. Further research in this field will be focused on the specification of other diagnostic indicators that will extend the possibilities of the comparison between the actual turbo-generator operation and nominal state based on the identified blade components in an even more sophisticated way. Further research could be focused on how to evaluate the phase of both spectral components of the blades in addition to amplitude and frequency. This could even help to localize where the blade excitation is applied around the blade wheel.

Author Contributions: Conceptualization, J.L., J.J. and V.V.; methodology, J.J. and V.V.; software, J.J. and V.V.; validation, J.J. and V.V.; formal analysis, J.J.; investigation, J.L., J.J. and V.V.; resources, J.L., J.J. and V.V.; data curation, J.J. and V.V.; writing—original draft preparation, V.V.; writing—review and editing, V.V., J.L. and J.J.; visualization, V.V.; supervision, J.L.; project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by ERDF under project "Research Cooperation for Higher Efficiency and Reliability of Blade Machines (LoStr)", No. CZ.02.1.01/0.0/0.0/16_026/0008389.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Russhard, P. The Rise and Fall of the Rotor Blade Strain Gauge. In *Vibration Engineering and Technology of Machinery*; University of Manchester: Manchester, UK, 2014; pp. 27–37. [CrossRef]
2. Heath, S.; Imregun, M. A Survey of Blade Tip-Timing Measurement Techniques for Turbomachinery Vibration. *J. Eng. Gas Turbines Power* **1998**, *120*, 784–791. [CrossRef]
3. Mathioudakis, K.; Loukis, S.; Papiliou, K. *Casing Vibration and Gas Turbine Operating Conditions*; American Society of Mechanical Engineers (ASME): Toronto, ON, Canada, 1989. [CrossRef]
4. Mathioudakis, K.; Papathanasiou, A.; Loukis, E.; Papiliou, K. Fast response wall pressure measurement as a means of gas turbine blade fault identification. *J. Eng. Gas Turbines Power* **1991**, *113*, 269–275. [CrossRef]
5. Gubran, A.A.; Sinha, J.K. Shaft instantaneous angular speed for blade vibration in rotating machine. *Mech. Syst. Sig. Process.* **2014**, *44*, 47–59. [CrossRef]
6. Bachschmid, N.; Salvini, G.; Tanzi, E.; Pesatori, E. The Influence of Blade Row Dynamics on Lateral and Torsional Shaft Vibrations in Steam Turbines. In Proceedings of the 9th IFToMM Intarnational Conference on Rotor Dynamics; Springer: Cham, Switzerland, 2015. [CrossRef]
7. Vasicek, V.; Liska, J.; Strnad, J.; Jakl, J. Identification of dynamic behavior of steam turbine blades using rotor vibration measurement. In Proceedings of the 14th European Conference on Turbomachinery Fluid dynamics & Thermodynamic, ETC 2021, Gdansk, Poland, 12–16 April 2021.
8. Vasicek, V.; Liska, J.; Strnad, J.; Jakl, J. Experimental validation of the blade excitation in a shaft vibration signals. In *Advances in Condition Monitoring and Structural Health Monitoring*; Springer: Singapore, 2021. [CrossRef]
9. Liska, J.; Jakl, J.; Vasicek, V. Rotating blades monitoring using standard turbine instrumentation. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2019**, *233*, 7447–7458. [CrossRef]
10. Liska, J.; Vasicek, V.; Jakl, J. On possibilities of using relative shaft vibration signals for rotating blades monitoring. In Proceedings of the Turbomachinery Technical Conference and Exposition, Oslo, Norway, 11–15 June 2018. [CrossRef]
11. Randall, R.B. *Vibration-Based Condition Monitoring*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2011. [CrossRef]
12. Vaseghi, S.V. *Advanced Digital Signal Processing and Noise Reduction*; John Wiley & Sons: Hoboken, NJ, USA, 2000; ISBN 0-470-84162-1. [CrossRef]

13. Proakis, J.G.; Manolakis, D.G. *Digital Signal Processing, Principles, Algorithms, and Applications*; Prentice-Hall: Hoboken, NJ, USA, 1996. [CrossRef]
14. Prabhu, K.M.M. *Window Functions and Their Applications in Signal Processing*; Taylor & Francis Group: Abingdon-on-Thames, UK, 2014. [CrossRef]
15. Deller, J.R.; Hansen, J.H.R.; Proakis, J.G. *Discrete-Time Processing of Speech Signals*; Willey-IEEE Press: Hoboken, NJ, USA, 1999; ISBN 978-0-780-35386-2.
16. Everitt, B.S.; Landau, S.; Leese, M.; Stahl, D. *Cluster Analysis*, 5th ed.; John Wiley & Sons Ltd.: Hoboken, NJ, USA; King's College: London, UK, 2011. [CrossRef]
17. Boyce, M.P. *Gas Turbine Engineering Handbook, 4th ed*; Elsevier: Amsterdam, The Netherlands, 2011. [CrossRef]

sensors

MDPI

Article

A Robust Deep Neural Network for Rolling Element Fault Diagnosis under Various Operating and Noisy Conditions

Chun-Yao Lee [1,*], Guang-Lin Zhuo [1] and Truong-An Le [2]

[1] Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan 320314, Taiwan; s10528245@cycu.org.tw
[2] Department of Electrical and Electronic Engineering, Thu Dau Mot University, Thu Dau Mot 75000, Binh Duong, Vietnam; anlt@tdmu.edu.vn
* Correspondence: cyl@cycu.edu.tw

Abstract: This study proposes a new intelligent diagnostic method for bearing faults in rotating machinery. The method uses a combination of nonlinear mode decomposition based on the improved fast kurtogram, gramian angular field, and convolutional neural network to detect the bearing state of rotating machinery. The nonlinear mode decomposition based on the improved fast kurtogram inherits the advantages of the original algorithm while improving the computational efficiency and signal-to-noise ratio. The gramian angular field can construct a two-dimensional image without destroying the time relationship of the signal. Therefore, the proposed method can perform fault diagnosis on rotating machinery under complex operating conditions. The proposed method is verified on the Paderborn dataset under heavy noise and multiple operating conditions to evaluate its effectiveness. Experimental results show that the proposed model outperforms wavelet denoising and the traditional adaptive decomposition method. The proposed model achieves over 99.6% accuracy in all four operating conditions provided by this dataset, and 93.8% accuracy in a strong noise environment with a signal-to-noise ratio of −4 dB.

Keywords: intelligent diagnostic; bearing faults; improved fast kurtogram (IFK); nonlinear mode decomposition (NMD); gramian angular field (GAF); convolutional neural network (CNN)

Citation: Lee, C.-Y.; Zhuo, G.-L.; Le, T.-A. A Robust Deep Neural Network for Rolling Element Fault Diagnosis under Various Operating and Noisy Conditions. *Sensors* **2022**, *22*, 4705. https://doi.org/10.3390/s22134705

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 8 May 2022
Accepted: 21 June 2022
Published: 22 June 2022

1. Introduction

As one of the basic components of modern industry, rotating machinery must ensure the reliability of its operation, among which bearings are an important part to maintain stability [1]. An IEEE study shows that bearing in induction machines fails most frequently, accounting for 42% of the total [2]. The early detection of failures to reduce maintenance costs and prevent unplanned downtime is a top priority for operators [3]. Therefore, it is an urgent task to develop a diagnostic system that can identify fault signals of motor bearing as early as possible.

With the continuous development of Big Data technology, industrial systems collect a large amount of operating data through sensors. How to effectively use these data is a major challenge for diagnostic methods [4]. Therefore, data-driven diagnostic methods have been proven to effectively utilize these signals to achieve accurate fault diagnosis [5]. Numerous studies on fault diagnosis have been published in recent years [6]. Song et al. proposed a signal analysis method combining statistical filter and wavelet packet transform with moving peak hold method, which can effectively extract weak fault signals of low-speed machinery [7]. Special bearing diagnostic symptom parameters are defined to extract sensitive features from the frequency domain. Van et al. combined non-local mean denoising and empirical mode decomposition (EMD) to accurately extract fault features [8]. The two-stage feature selection of hybrid distance evaluation technology (DET) and particle swarm optimization (PSO) can effectively divide the feature interval and find the best feature subset. Wang et al. proposed a model that integrates fault diagnosis

and prediction [9]. The model adopts the multi-scale envelope spectrum to analyze the fault characteristics and expresses it as a bearing health index, and then estimates the residual life by Bayesian inference. Despite the success of the above studies, the main limitations of these methods are that the feature extraction process is highly dependent on expert experience and experiments, and the diagnostic accuracy is highly dependent on the quality of the selected features [10]. The process of manually selecting features is usually time-consuming and only suitable for specific tasks [11]. In addition, the fully connected structure of feature extraction, feature selection, and classifier is not deep enough, which is a challenge for the classification task of complex systems [12]. Therefore, deep learning (DL), which can automatically complete feature extraction and has a deep structure capable of learning the complex relationship between signals and features, is the current research trend [13].

Recently, a large number of bearing fault diagnosis studies using DL have been proposed [14]. Among them, the convolutional neural network (CNN) is particularly suitable for developing bearing intelligent diagnosis models due to its sparse connection, weight sharing, and performance in processing periodic signals [15,16]. Huang et al. proposed a multi-scale cascaded CNN (MC-CNN) to find useful frequency bands through filters of different scales to enhance the input information [17]. Li et al. proposed a method for intelligent bearing remaining useful life prediction using a short-time Fourier transform (STFT)-based time–frequency map combined with CNN [18]. Darong et al. proposed to combine a modified recursive least squares (RLS) model based on the momentum factor and the forgotten factor with local mean decomposition (LMD) for early bearing fault diagnosis [19]. Zhao et al. proposed a deep rational attention network (DRANet) for fault diagnosis. In this method, signal denoising is introduced and the proposed pseudo-soft threshold function is used to avoid gradient vanishing [20]. However, the robustness of these studies to noise and variable operating conditions is still insufficient. Because motors often operate at variable speeds and noisy environments in industrial environments, these strong interference components mask the fault pulses, making the traditional CNN-based fault diagnosis models have poor fault diagnosis performance under variable conditions. Tang et al. propose a multiscale CNN that integrates a vision transformer (ViT) and continuous wavelet transform (CWT). The model integrating CWT and ViT can offer more hidden fault-related information from multi-scale components and achieve higher generalization and anti-noise performance [21]. Qiao et al. proposed an adaptive weighted multiscale convolutional neural network (AWMSCNN) to address the domain shift problem that may be caused by fault diagnosis under variable working conditions. The AWMSCNN with several convolution kernels of different widths and adaptive weight vectors has strong fault discrimination ability and domain adaptation ability under variable working conditions [22]. Qin et al. proposed a deep twin convolutional neural networks with multi-domain inputs (DTCNNMI), which integrated time–domain, time–frequency domain, and time–domain statistical features. DTCNNMI was successfully validated under strong noise and different operating conditions [23]. Qiao et al. proposed a dual input of time–domain signals and time–frequency images combined with CNN and long short-term memory (LSTM) to study fault diagnosis under variable loads and different noise conditions [24]. Jin et al. proposed a CNN with an attention mechanism and adopted a random sampling strategy and an exponential linear unit as the activation function to improve the adaptability of the network under complex conditions [25]. Zhang et al. proposed a CNN with wide first-layer kernels to extract features and suppress high-frequency noise using wide kernels in the first convolutional layer, namely WDCNN [26].

Based on the above review, although many diagnostic models can achieve good classification results, not all diagnostic models can achieve high-precision diagnosis in complex environments, especially under noisy and changing load conditions, and there are still relatively few studies in this part. Ensuring classification efficiency in complex environments is the goal of this study. Therefore, this study introduces a mono-component decomposition method with strong noise immunity, called nonlinear mode decomposition

(NMD) [27]. NMD integrates time–frequency analysis [28], surrogate data testing [29], and harmonic identification [30]. Research on the fault diagnosis of rotating machinery [31,32] has demonstrated its noise robustness and only extracts physically meaningful oscillations. Moreover, an improved fast kurtogram (IFK) is proposed to address the computational inefficiency of NMD pointed out by [32]. Furthermore, the classification performance of CNN is highly dependent on dataset quality. Therefore, the gramian angular field (GAF) is introduced to construct a 2D image, which can obtain a unique temporal correlation mapping in polar coordinates [33]. In summary, this paper proposes an intelligent fault diagnosis model, called IFKNMD-CNN, which uses CNN to classify a dataset constructed by an innovative combination of IFK, NMD, and GAF. The advantages and contributions of this paper are summarized as follows.

1. The combination of IFK and NMD not only greatly improves the computational efficiency, but also has a high tolerance to noise. Because IFK finds the best frequency band of the signal, it filters out redundant parts and noise in the signal.
2. NMD uses a surrogate data test to ensure that the output is a physically meaningful component rather than noise; therefore, the fault diagnosis model can achieve robust performance even in highly noisy environments.
3. GAF can obtain a unique time map in the polar coordinate system, fully demonstrating the advantages of CNN in classifying bearing signals.
4. This study uses the public dataset provided by Paderborn University to verify the effectiveness of the model [34]. The performance of the proposed model is validated in comparison with other state-of-the-art methods using the same dataset. Furthermore, the proposed model is tested under the four operating conditions provided by the Paderborn dataset, proving that the model also has a high degree of generalization.

The rest of this article is organized as follows. Section 2 presents the theoretical background of the proposed model. Section 3 details the performance improvement of the proposed method and the diagnostic steps of the proposed model. Section 4 discusses the experimental results on the Paderborn University dataset. Section 5 carefully evaluates the model. Finally, Section 6 summarizes the fault diagnosis model.

2. Background Theories

2.1. Nonlinear Mode Decomposition

The merits of NMD in adaptive mode decomposition and anti-noise capability are based on a powerful combination of time–frequency analysis, surrogate data test, and harmonic identification [27]. Detailed descriptions of the NMD are as follows:

Step 1: The time–frequency representation (TFR) of a given vibration signal $x(t)$ is calculated from the windowed Fourier transform (WFT) defined as $G_x(\omega,t)$. Next, the dominant component is extracted from the TFR, and its characteristic parameters (including instantaneous amplitude, instantaneous phase, and instantaneous frequency) are reconstructed by the ridge method. The corresponding formulas of the ridge method are as follows:

$$\omega_p(t) = \text{argmax}_\omega |G_x(\omega, t)| \tag{1}$$

$$\omega(t) = \omega_p(t) + \delta\omega_d(t) \tag{2}$$

$$A(t)e^{i\phi(t)} = \frac{2G_x\left(\omega_p(t), t\right)}{\hat{g}\left[\omega_p(t) - \omega(t)\right]} \tag{3}$$

$$x_d(t) = \text{Re}\left[A(t)e^{i\phi(t)}\right] \tag{4}$$

where $\omega_p(t)$ is the ridge curve of the dominant component, $\delta\omega_d(t)$ is the correction for discretization effects, the Gaussian window for the WFT is denoted as $\hat{g}(w) = e^{-(f_0 w)^2/2}$, and $x_d(t)$ is the dominant component to be extracted.

Step 2: The Fourier transform (FT) surrogate test can effectively identify the reference component or noise. This method is constructed by taking the inverse Fourier transform of the signal's FT and randomizing the phase of the Fourier coefficients:

$$x_x(t) = (2\pi)^{-1} \int \left[\hat{x}(\xi) e^{i\phi_x(\xi)}\right] e^{i\xi t} d\xi \tag{5}$$

where $\phi_x(\xi)$ represents a uniformly random phase taken on [0, 2π]. If the reference component is true, then it should be more deterministic, since randomization will destroy the phase relationship of the amplitude and frequency modulation of the surrogate component, making it less deterministic.

Spectral entropy quantifies the degree of determinacy in the extracted amplitude $A(t)$ and frequency $\omega(t)$, and the combination of the spectral entropies is the discriminant statistic for the surrogate test. The formulas of spectral entropy and discriminant statistics D are defined as:

$$Q[f(x)] \equiv -\int \frac{|f(x)|^2}{\int |f(x)|^2 dx} \log \frac{|f(x)|^2}{\int |f(x)|^2 dx} dx \tag{6}$$

$$D(\alpha_A, \alpha_\omega) \equiv \alpha_A Q\left[\hat{A}(\xi)\right] + \alpha_\omega Q[\hat{\omega}(\xi)] \tag{7}$$

The significance D_s is defined as the maximum value among D(1,0), D(0,1), and D(1,1). D_0 is the significance of the original component. In this paper, N_s = 40 FT surrogates are built, and the significance level is set to λ = 95%. The corresponding significance $Ds = 1, 2, \ldots, N_s(\alpha_A,\alpha_\omega)$ is calculated and compared with the original components. If the surrogates for $D_s > D_0$ are not less than λN_s, the original component is considered true.

Step 3: The time–shift surrogate test is used to check whether the dominant component $x_d(t)$ can represent the first harmonic (fundamental). The time-shifted surrogate method can build harmonic surrogates consistent with the null hypothesis of independence and reduce the interference caused by noise, finite frequency, and time resolutions. In the time–shift surrogate test, $x_d(t)$ is assumed to be the fundamental harmonic $x_1(t)$. Next, the corresponding candidate harmonics for $i = 1/2, 1/3, \ldots$ are extracted from the time-shifted TFR. The formula for the instantaneous amplitude $A_i(t)$ of the subharmonic $x_i(t)$ is constructed by shifting $\Delta T_d/2$ backward:

$$A_d^i(\tau) = A^i(\tau - \Delta T_d/2) \tag{8}$$

$$\tau = t_{i=1+M/2,\ldots,N-M/2} \tag{9}$$

$$\Delta T_{d=1,\ldots,N_d} = M(1 - 2d/N_d)/2f_s \tag{10}$$

where N is the total length of the subharmonic, M is the maximal time–shift, d represents the index of candidate harmonics, N_d indicates the number of candidate harmonics, and f_s is the sampling frequency of signal.

A metric value $q_A^i \in [0, 1]$ is designed to quantify the degree of dependence between the first harmonic and the extracted harmonic candidates (0: no consistency, 1: full consistency). A metric value of amplitude $q_A^i \in [0, 1]$ is defined as:

$$q_A^i \equiv \exp\left[-\frac{\sqrt{\left\langle \left[A^i(t)\langle A^1(t)\rangle - A^1(t)\langle A^i(t)\rangle\right]^2\right\rangle}}{\left\langle A^1(t)A^i(t)\right\rangle}\right] \tag{11}$$

The overall metric of interdependence between the harmonics is defined as:

$$\rho^i = (q_A^i)^{\alpha_A} \tag{12}$$

where α_A is the weights of each metric q_A^i. According to (8)–(11), the consistency of the candidate harmonics $\rho^i_{d=1,\ldots,N_d}(1)$ is calculated and compared with the value $\rho_0^i(1)$ of the

zero time shift $\Delta T_0 = 0$ for consistency comparison. The ratio of $\rho^i_{d=1,\ldots,N_d} > \rho^i_0$ represents the probability that the ith harmonic is a true harmonic. In this paper, the probability is set to be $\geq$95% and the number of candidate harmonics N_d is set to 100. In addition, to reduce the probability of false positives caused by noise, a threshold $\rho_{\min} = 0.25$ is set in the time-shifted surrogate test. The harmonics are considered true only if they pass the surrogate test while being greater than the threshold.

Step 4: The true harmonic with the smallest index i found in the previous step is the reference component for extracting high-order harmonics. *Step 3* is repeated and the high-order harmonics are stored as $i = 2, 3, \ldots$.

Step 5: The true harmonics are subtracted from the given signal $x(t)$ and all the above steps are repeated until the stopping condition is met, i.e., the residual is identified as noise.

NMD is a powerful combination based on time–frequency analysis, surrogate data testing, and harmonic identification. NMD extracts dominant components through time–frequency analysis, FT surrogate data tests, and time–shift surrogate data tests, removing interfering components such as noise. A metric is then defined to identify physically meaningful modes. Therefore, NMD achieves adaptive decomposition and robustness to noise. In this study, the instantaneous amplitude in the mode can be extracted and used to construct images further that clearly reflect the fault features of the test motor.

2.2. Gramian Angular Field

GAF is used to map a time series signal to a polar coordinate system, encoding the signal into a unique time map to help CNNs perform high-precision classification [10]. The time series $X = \{x_1, x_2, \ldots, x_n\}$ with n samples is normalized so that all values fall within the interval $[-1, 1]$, as in (13). The normalized time series $\tilde{X}$ is concerted into a polar coordinate system map, with values encoded as the angular cosine (bounded by $[0, \pi]$) and timestamps as the radius. Note that a given time series produces one and only one map in the polar coordinates system, with a unique inverse map, as in (14). Finally, the trigonometric sum/difference between each sample is calculated to construct an image that preserves the absolute temporal relationship, as in (15) and (16).

$$\tilde{x}_i = \frac{(x_i - \max(X)) + (x_i - \min(X))}{\max(X) - \min(X)} \tag{13}$$

$$\begin{cases} \phi_i = \cos^{-1}(\tilde{x}_i), & -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ \gamma = t_i / N, & t_i \in N \end{cases} \tag{14}$$

$$\text{GASF} = \left[\cos(\phi_i + \phi_j)\right] \tag{15}$$

$$\text{GADF} = \left[\sin(\phi_i - \phi_j)\right] \tag{16}$$

where i is the length of the given signal, t_i is the timestamp, and N is a constant used to regularize the span of the polar coordinate system. GAF has the following several advantages: (1) according to (15) and (16), the temporal correlation is represented by superposition/difference for time intervals; (2) in the GAF matrix, the diagonal of the value contains the original value and angular information. Based on the above advantages, GAF can provide high-quality images for CNN to learn the complex relationship between the various health states of the motor. Figure 1 shows the transformation of time series signals into GAF images. Vibration signals from the same bearing dataset are used as experimental results in Section 4. Important features in the signal are highlighted (red part). Moreover, it can be observed from Figure 1 that the important feature distributions of the three states are significantly different, indicating that GAF can clearly express the complex information of rotating machinery.

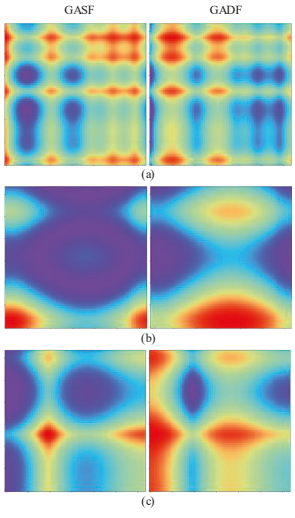


Figure 1. Illustration of GAF image: (**a**) healthy; (**b**) outer ring fault; and (**c**) inner ring fault.

2.3. Convolutional Neural Network

A typical CNN contains numerous modules consisting of convolutional and pooling layers, followed by a fully connected layer. The CNN structure used in this paper contains three convolutional pooling modules, a fully connected layer, and an output layer. The detailed description of the convolution pooling module and the fully connected layer is as follows.

(1) *Convolutional Pooling Module:* In the convolutional layer, a set of convolution kernels consisting of weights and biases performs convolution operations on the input image with a specific stride. Because the same kernel is used to extract features during the convolution operation, the number of neural network parameters can be greatly reduced, and the operation efficiency can be improved. This advantage is called weight sharing. The feature maps are obtained after the convolution operation changes with the weights of the convolution kernel. The activation function is used to perform a non-linear transformation on the feature maps. This operation can help improve model performance. It can be assumed that the convolutional layer l has D convolution kernels, and $n = 1, 2, \ldots, D$. The output of the nth convolution kernel c_n^l can be expressed as:

$$c_n^l = \mathrm{ReLU}(\sum_j I_j^{l-1} \odot w_n^l + b_n^l) \tag{17}$$

where I_j^{l-1} is the jth output of the previous layer $l-1$; w_n^l and b_n^l are the weights and bias of the nth convolutional kernel in convolutional layer l; $\odot$ indicates the convolution operation; and ReLU represents the activation function.

The pooling layer is usually set after the convolutional layer. The purpose is to down-sample the output of the convolutional layer. The advantage of this operation is to reduce the feature dimension of the output without losing features, which can further improve training efficiency and reduce memory usage. The definition of the max-pooling operation adopted in this study is as follows:

$$P_n^{l+1} = \max_{x \times y}(c_n^l) \tag{18}$$

where P_n^{l+1} is the output of the max-pooling operation; x and y represent the size of the pooling region; and the size of the pooling region is set to 2×2 in this study.

(2) *Fully Connected Layer:* The fully connected layer can integrate the local features of the motor state obtained in the convolution pooling module and pass it to the output layer for classification. In the fully connected layer, the learned features are flattened into a one-dimensional feature vector, and the classification results are obtained by adjusting the weights and biases. To improve CNN performance, each neuron in the fully connected layer uses an activation function.

3. Proposed Method

3.1. Speed Acceleration for the NMD Based on IFK

Although NMD is a powerful adaptive decomposition algorithm, the length and frequency range of the signal can seriously affect the computational efficiency of NMD. Therefore, this paper proposes an improved FK (IFK) to find the optimal filtering band of the signal and greatly improve the computational efficiency of NMD. A detailed description of the traditional FK can be found at [35]. In addition, the NMD is limited to searching for candidate components in the optimal filtering frequency band. The IFK greatly improves the diagnostic accuracy using the comprehensive index of the clearance factor and kurtosis. The IFK has three advantages: (1) it resolves any problems after the traditional FK method incorrectly selects the frequency band due to the influence of any other pulse signals in the environment; (2) for rotating machinery, the clearance factor can better separate healthy bearings from faulty bearings; (3) the IFK can extract the fault signal covered by noise and present it as an envelope signal, which can further improve the signal-to-noise ratio (SNR) from the motor vibration signal. The detailed description of IFK is shown in Algorithm 1.

Algorithm 1: Improved FK.

Input: the vibration signal x
Output: the complex envelope x_{ce} positioned on the central frequency f
1: predefine 1/3 binary tree filter banks; low-pass and high-pass analysis filters $h_0(n)$ and $h_1(n)$
2: **for** $k = 0$ to L-1 **do**
3: define $c_k^i(n)$ as the sequence of coefficients obtained from the ith filter // where $i = 0, \ldots, 2^k - 1$
4: **if** $k = 0$: $c_0(n) \equiv x(n)$
5: calculate two new coefficients $c_{k+1}^{2i}(n)$ and $c_{k+1}^{2i+1}(n)$ by $h_0(n)$ and $h_1(n)$
6: calculate the kurtosis K_k^i and clearance factor C_k^i of all coefficients // where $i = 0, \ldots, 2^k - 1$
7: **end for** ***k***
8: calculate the comprehensive index of each coefficient Q_k^i
9: obtain the x_{ce} based on the coefficient which has the best Q

The $h_0(n)$ and $h_1(n)$ can be represented as $h_0(n) = h(n)e^{j\pi n/4}$ and $h_1(n) = h(n)e^{j3\pi n/4}$, respectively; the Q_k^i is the mean of the statistic values of K_k^i and C_k^i.

A comparison study of the improved computational efficiency of the NMD method and the original version is carried out. The vibration signal from the same bearing dataset is used as the experimental results in Section 4, and the sampling rate of the signal is 64 kHz. In total, 4 seconds of vibration signal is measured, so the signal length reaches 256,000 data points. In this dataset, bearing codes K, KA, and KI denote healthy bearing, outer ring fault, and inner ring fault, respectively. NMD is limited to searching for candidate components in the optimal filtering band indicated by the IFK. Figure 2 shows the results of selecting the optimal frequency band for the vibration signal of a bearing outer ring fault (KA30). The results show that IFK finds a narrower frequency band. In contrast, FK cannot accurately separate fault pulses or interfering pulses in the environment. In addition, the computational efficiency of the original FK is also compared. The comparison results are shown in Table 1. The original version of the NMD method requires a large amount of time for mono-component decomposition. However, the mono-component extraction for the three bearing states can be accomplished within 22 seconds by the improved NMD method. Moreover, the comparison results with the original FK show that the IFK can find a narrower filtering band to achieve computational efficiency improvements and make it

suitable for real-world cases. This comparison study is conducted with an Intel(R) Core (TM) i5-10500 3.1 GHz CPU and 16.0 GB RAM.

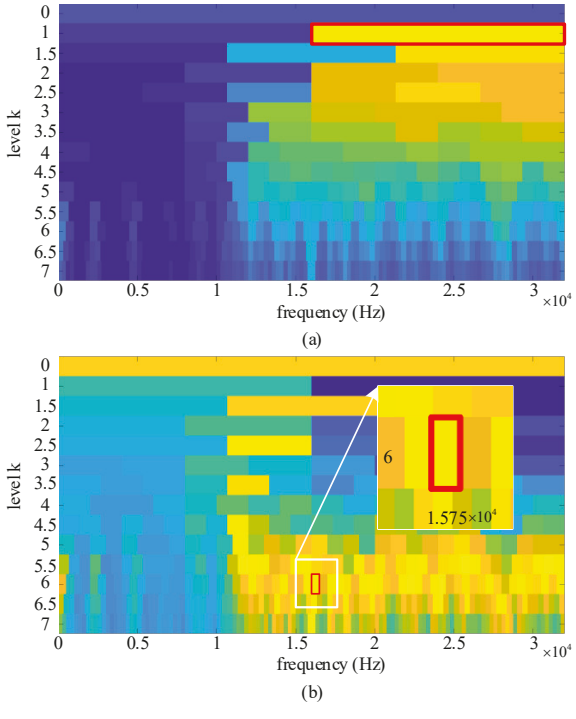


Figure 2. The best frequency band selected by (**a**) FK and (**b**) IFK.

Table 1. Comparison of computational time.

Bearing Code	Original Signal	FK			IFK		
	t_c (s)	Level	f_c (Hz)	t_c (s)	Level	f_c (Hz)	t_c (s)
K004	3125	1	24,000	543	4.5	19,333	**22**
KA22	8691	1	24,000	879	6	18,250	**12**
KI14	7322	3.5	30,666	43	6.5	30,167	**10**

The f_c and t_c indicate central frequency and computational time, respectively. At each level, the signal length of the filtered sequence is reduced by a factor of 2. Therefore, the length of the sequence obtained at level 1 is 127,985, the length of the sequence obtained at level 2 is 63,984, and so on.

3.2. Proposed Rolling Element Fault Diagnosis Model

The proposed induction motor rolling element fault diagnosis model, namely IFKNMD-CNN, is shown in Figure 3. In this model, the vibration signal is obtained from the test rig. Then, the IFK is used to find the best filtering frequency band from the vibration signal and extract the envelope signal. Next, NMD extracts physically meaningful components from the envelope signal and uses GAF to perform 2-D image transformation on the components afterward. Finally, CNN is used for rolling element fault classification. The detailed process of IFKNMD-CNN is summarized as follows.

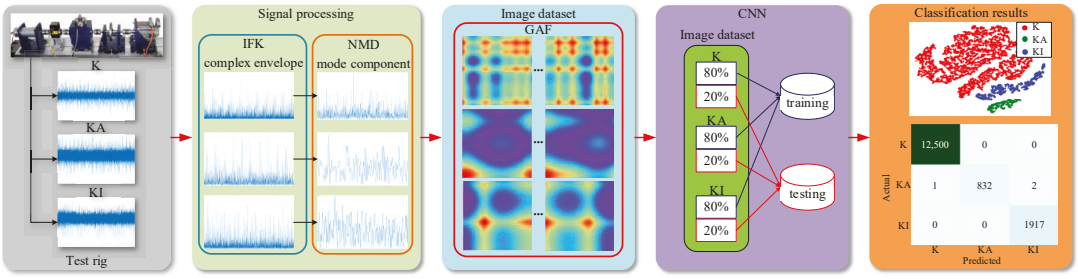


Figure 3. Illustration of a fault diagnosis model.

Step 1: The accelerometer is used to collect three kinds of vibration signals from the test platform, including the health, inner ring fault, and outer ring fault.

Step 2: The IFK is proposed to find the narrow frequency band containing the main fault information in the vibration signal, filter out the part outside the narrow frequency band through low-pass and high-pass filters, and present the reserved part as an envelope signal, which can greatly reduce the signal length and improve the signal SNR.

Step 3: The NMD method is used to further analyze the envelope signal and extract the mode component containing fault characteristics. NMD is constrained to search for components within the optimal frequency band indicated by the IFK. The combination of the IFK and NMD greatly increases computational efficiency.

Step 4: GAF is used to transform the time domain signal into a polar coordinate system map and preserve the temporal correlation so that it can construct a high-quality 2D image dataset.

Step 5: CNN is adopted to learn the image dataset and perform fault classification. During the classification process, 80% of the samples for each health state of the motor in the image dataset are randomly selected as the training dataset and 20% as the test dataset. Randomly selecting samples can ensure that each classification process uses a different dataset, avoid possible overfitting, and prove the effectiveness of the fault diagnosis model.

4. Case Study

4.1. Experimental Setup

This study uses the Paderborn dataset proposed by Lessmeier et al. This dataset contains both healthy bearings and widely distributed damage of inner and outer ring bearings. Real damaged bearings are produced by an accelerated lifetime test rig. The damage produced by the lifetime test mainly occurs in different degrees of pitting. Indentations are found on a small number of bearings. A detailed description of the real damaged bearing can be found in [34]. Then, the test bearings are mounted to the modular test rig to generate experimental data. The test rig uses a piezoelectric accelerometer (Model No. 336C04, PCB Piezotronics, Inc., Depew, NY, USA) with a sampling rate of 64 kHz to measure the vibration signal at the adapter at the top end of the rolling bearing module. Each measurement has a duration of 4 seconds and is repeated 20 times independently. The detailed parameter settings of the Paderborn dataset used in this study are shown in Table 2. In this study, data with a length of 128,000 are selected from each fault category for signal analysis (the size of the 2-D image is set to 64×64).

Table 2. Parameter setting of the experimental dataset.

Rotational Speed [rpm]	Load Torque [Nm]	Radial Force [N]	Name of Dataset
1500	0.7	1000	N15_M07_F10
900	0.7	1000	N09_M07_F10
1500	0.1	1000	N15_M01_F10
1500	0.7	400	N15_M07_F04

4.2. Image Dataset for DL Method

Four DL models are compared in this case study, including IFKNMD-CNN, CNN based on LMD with TFR (LMD-TFR-CNN), IFKNMD-1DCNN, and 1D-CNN. In LMD-TFR-CNN, LMD can adaptively decompose the signal into a set of product functions (PFs), and then use PF selection [36] to select the best PF and express it as a time–frequency relationship. IFKNMD-1DCNN is an ablation version of the proposed model, which skips GAF and directly uses one-dimensional signal as the input of CNN. In addition to IFKNMD-CNN and IFKNMD-1DCNN, the data used by both LMD-TFR-CNN and 1D-CNN introduce a wavelet denoising technique to verify the anti-noise capability of the proposed model. In this technique, the selected signal is decomposed using a multi-level decomposition parameterized as a wavelet function (db4) and a hard threshold, and then the signal is reconstructed using an inverse wavelet transform. In this subsection, an operating state, namely the N15_M07_F10 dataset, is used as an example to demonstrate the experimental results.

(1) *Image Dataset for IFKNMD-CNN:* As described in the previous section, the IFK is used to find the optimal frequency band of the signal and represent it as an envelope signal. Table 3 shows the analytical results of the IFK for the bearing categories used in this experiment. The health state (K) has a lower decomposition level, and the faulty state (KA and KI) has a higher decomposition level. This result shows that the IFK can accurately capture the frequency band where the main fault pulses are concentrated, only maintaining the important part of the signal for fault diagnosis. Moreover, the length of the original signal used in this case study is 128,000 and the average length of the extracted envelope signal (15 categories) is 14,712. More than 88% of the unimportant parts of the signal are removed. As verified by the comparison experiments in Section 3, the computational burden of the model in the signal analysis is greatly reduced.

Table 3. Analysis results of IFK.

Bearing Code	Level	f_c (Hz)	Signal Length
K001	1.5	16,000	42,642
K002	3	26,000	15,984
K003	1	24,000	63,985
K004	1.5	16,000	42,642
K005	3	30,000	15,985
KA04	6	30,750	1985
KA15	7	17,375	985
KA16	6	9250	1984
KA22	6.5	16,167	1309
KA30	5.5	24,333	2642
KI04	5.5	14,333	2642
KI14	6.5	24,500	1309
KI16	4.5	18,000	5309
KI18	3.5	28,000	10,642
KI21	3.5	12,000	10,642

The instantaneous amplitudes obtained from NMD are divided into segments of length 6400. As shown in Table 3, segments with a length of 6400 are kept, and segments

less than 6400 are discarded, such as all health state data, KI18, and KI21. If the signal length is less than 6400, it will not be divided, including all outer ring fault data, KI04, KI14, and KI16. Then, each segment is transformed into an image using GAF. If the segment length exceeds 3200, the transformed image size is 3200 × 3200. An image of this size is sufficient to reflect the motor state information without distortion. For all outer ring fault data, KI04, and KI14, the transformed image size is the signal length × signal length. To construct an image dataset of sufficient size to train the CNN, each image is divided into 64 × 64 size images. As shown in Table 4, the dataset contains 76,248 samples. Healthy samples accounted for 81.9% of the total samples. Highly imbalanced datasets are also more suitable for actual applications.

Table 4. Details of N15_M07_F10 image dataset for IFKNMD-CNN.

Bearing Code	No. Sample	Total Samples	Percent
K001	15,000	62,500	81.9%
K002	5000		
K003	22,500		
K004	15,000		
K005	5000		
KA04	961	4167	5.5%
KA15	225		
KA16	900		
KA22	400		
KA30	1681		
KI04	1681	9581	12.6%
KI14	400		
KI16	2500		
KI18	2500		
KI21	2500		

(2) *Image Dataset for LMD-TFR-CNN:* For a fair comparison, the selected best PF is also divided into segments of length 6400, resulting in a total of 20 segments. Each segment is transformed into a TFR (297 × 6400) according to a continuous wavelet transform-based time–frequency analysis. Similarly, each time–frequency image is divided into small 64 × 64 images. Therefore, the dataset has a total of 120,000 samples.

(3) *Image Dataset for IFKNMD-1DCNN:* The instantaneous amplitude obtained from the NMD is divided into segments of length 64. As shown in Table 3, segments whose length is less than 64 are discarded. Therefore, there are 2829 healthy samples, 138 bearing outer ring failure samples, and 475 inner ring failure samples, for a total of 3442 samples.

(4) *Image Dataset for 1D-CNN:* The new signal after wavelet denoising is divided into segments of length 1024, and a total of 125 segments are obtained. There are 15 categories in the bearing dataset, so the dataset has a total of 1875 samples.

4.3. Parameters Setting of CNN

The images obtained by IFKNMD-CNN (IFK, NMD, and GAF) can clearly reflect the fault features, and then CNN is used to complete high-accuracy fault diagnosis. The structure of the CNN network is shown in Figure 4. The input image size is 64 × 64. The convolution kernels of the three convolutional layers have the same size (3 × 3) and the numbers are 16, 32, and 32, respectively. Padding is set to (1, 1). The activation function adopts ReLU. The number of neurons in the fully connected layer is 512. The output size is 3, corresponding to the number of categories for this case study. The number of training epochs is 500. The Adam algorithm is adopted in the training process with a learning rate of 0.001. This experiment is performed on a computer with Intel(R) Core(TM) i5-10500 3.1 GHz CPU and GEFORCE GTX 1050 GPU running PyTorch 1.10.0.

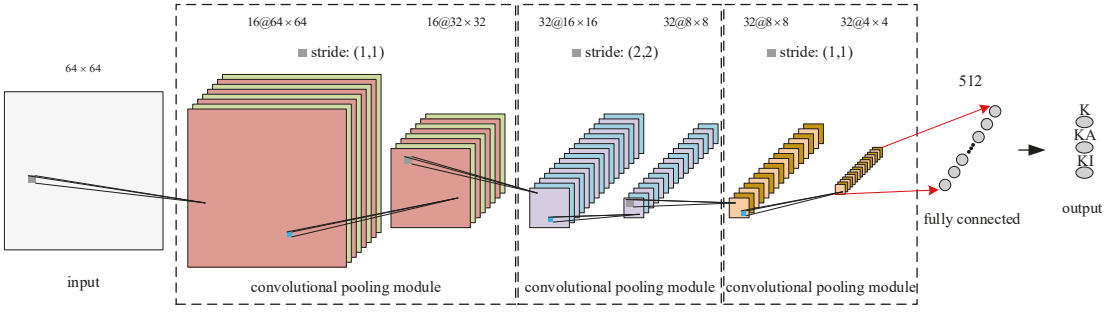


Figure 4. Structure of CNN network.

4.4. Intelligent Diagnosis with IFKNMD-CNN

(1) *Performance Analysis of IFKNMD-CNN:* T-SNE and confusion matrix are used in this case study to validate the performance of the proposed model. *Accuracy* and *F*-score are metrics for evaluating model performance, as defined by the following equations:

$$accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i} \tag{19}$$

$$precision = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{20}$$

$$recall = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{21}$$

$$F = \frac{2 \times precision \times recall}{precision + recall} \tag{22}$$

where n represents the three bearing states, and TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively.

The visualization results of the features of the fully connected layer in the CNN are shown in Figure 5. The features of IFKNMD-CNN (see Figure 5a) are clearly separated. This shows that IFKNMD-CNN can extract useful features to help CNN perform high-accuracy fault classification. In contrast, IFKNMD-1DCNN can distinguish most of the healthy features but cannot effectively distinguish the fault states (KA and KI) (see Figure 5b). The classification results can be more clearly observed through the confusion matrix in Figure 6. Both IFKNMD-CNN and IFKNMD-1DCNN achieve 100% accuracy in recognizing the health state. This indicates that the signals processed by IFK and NMD can clearly distinguish the healthy state. IFKNMD-CNN can separate inner ring fault and outer ring fault more effectively than IFKNMD-1DCNN. In this case study, the accuracy of IFKNMD-CNN (99.96%) is higher than that of IFKNMD-1DCNN (98.41%). This shows that the two-dimensional image obtained by GAF can provide more useful information than the one-dimensional signal. In addition, Figure 7 shows the receiver operating characteristic (ROC) curve as well as the area under the curve (AUC) of the DL method. IFKNMD-CNN achieves the highest accuracy (AUC = 1.00).

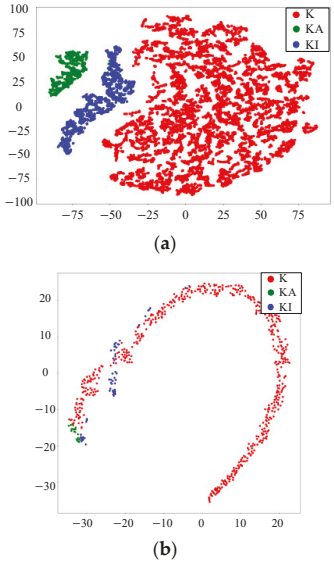


Figure 5. Visualization results with (**a**) IFKNMD-CNN and (**b**) IFKNMD-1DCNN.

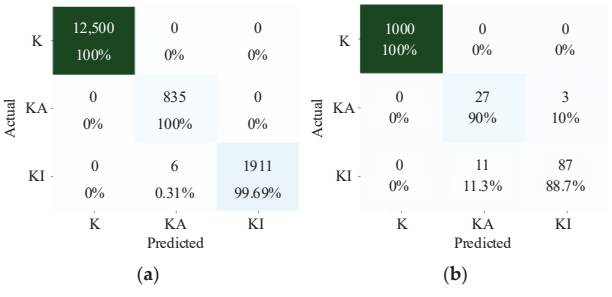


Figure 6. Confusion matrix with (**a**) IFKNMD-CNN and (**b**) IFKNMD-1DCNN.

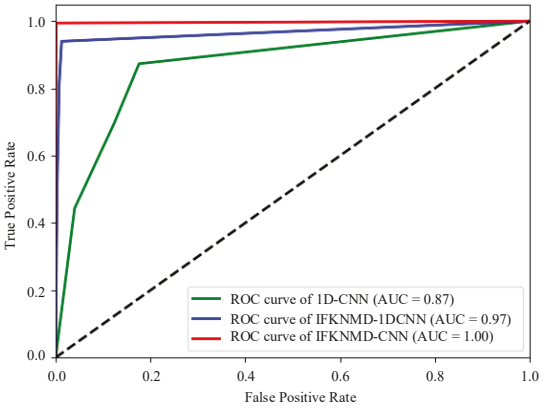


Figure 7. ROC curve with comparison method.

In addition, the robustness of IFKNMD-CNN to noise is also evaluated. In this experiment, white Gaussian noise (23) with SNR ranging from −4 to 10 dB is injected into the original signal to construct a new signal corresponding to the SNR.

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log 10\left(P_{\mathrm{signal}}/P_{\mathrm{noise}}\right) \tag{23}$$

As shown in Figure 8, IFKNMD-CNN and IFKNMD-1DCNN significantly outperform 1D-CNN, while IFKNMD-CNN achieves the best anti-noise capability. The results show that the combination of the optimal band filtering strategy of IFK and the mono-component decomposition of NMD has better anti-noise capability. IFKNMD-CNN achieves an average accuracy greater than 93%. Under the SNR of −4 dB, the accuracy of IFKNMD-1DCNN drops significantly, reaching 5.8%. The overall accuracy of 1D-CNN is lower than that of IFKNMD-CNN, and the fluctuation is obvious. When the SNR value is −4 dB, the accuracy difference between IFKNMD-CNN and 1D-CNN is the largest, reaching 28.8%. There are two factors that lead to these experimental results: (1) the feature similarity between bearing faults (KA and KI) is high, which increases the difficulty of classification; (2) the energy of bearing fault features is very weak, which is difficult to classify correctly in a strong noise environment.

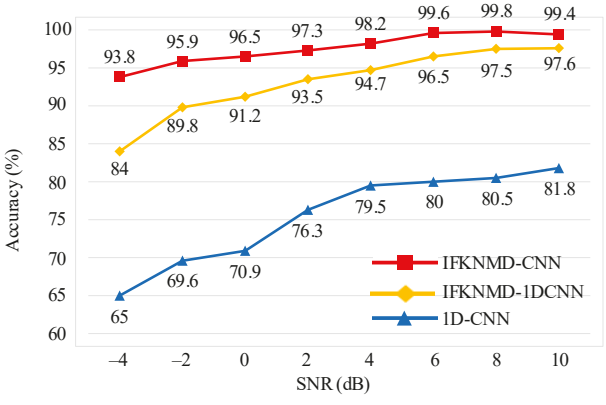


Figure 8. Noise test results with IFKNMD-CNN and comparison methods.

(2) *Comparison With DL Methods:* Table 5 shows the average accuracy and mean *F*-score for 10 independent runs of the four methods. IFKNMD-CNN can identify the health states with almost 100% accuracy. IFKNMD-1DCNN also performs well in identifying healthy states but suffers many classification errors in classifying faulty states (KA and KI). LMD-TFR-CNN introduces a PF function selection technique to remove unimportant PF functions [36]. However, it can be observed from the classification results that LMD is not sensitive to complex relationships between states, and cannot generate high-quality images for CNN to learn features. In addition, although LMD-TFR-CNN and 1D-CNN introduce wavelet denoising to improve the SNR, the wavelet transform is limited by the defined frequency band, which makes it non-adaptive. Although IFKNMD-1DCNN achieves accuracy of 98.62%, the *F*-score is only 87.37%. The main reason for this result is the high similarity between the features of bearing faults (KA and KI), which makes the model unable to classify effectively. The same results can also be observed in 1D-CNN. IFKNMD-CNN successfully highlights the important features of each state by mapping the signal to the polar coordinate system through GAF. In conclusion, IFKNMD-CNN achieves the highest accuracy and *F*-score, proving that the proposed model can achieve robust fault diagnosis.

Table 5. Comparison results with deep learning methods.

Method	*Accuracy* (%)				*F* (%)
	K	KA	KI	Overall	
IFKNMD-CNN	**100**	**98.3**	**99.35**	**99.82**	**99.31**
IFKNMD-1DCNN	99.95	91.58	76.11	98.62	87.37
LMD-TFR-CNN	88.58	97.27	93.67	93.17	93.38
1D-CNN	97.2	85.68	78.64	87.17	87.14

(3) *Apply to Various Operating Conditions:* This experiment validates the generalization ability of the fault diagnosis model by providing a variety of operating conditions through the bearing dataset. The test rig adjusts rotational speed, load torque, and radial force to construct data for four operating states. The experimental results are shown in Table 6. The accuracies in the four operating conditions are all >99.5%. Therefore, IFKNMD-CNN can be applied to a variety of operating conditions.

Table 6. Experimental results in different operating conditions.

Dataset	*Accuracy* (%)				*F* (%)
	K	KA	KI	Overall	
N15_M07_F10	100	98.3	99.35	99.82	99.31
N09_M07_F10	100	98.63	97.15	99.65	98.55
N15_M01_F10	100	98.38	96.74	99.6	98.41
N15_M07_F04	100	96.21	97.43	99.67	97.91

4.5. Comparison with Existing Methods

The authors of the Paderborn dataset proposed a fault diagnosis method that combines feature extraction, feature selection, and classifiers [34]. Time–domain features, frequency domain features, and time–frequency domain features are first extracted. Then, the maximum separation distance is applied to feature selection. Finally, different classifiers are used for fault classification. In this subsection, three classifiers with the highest classification accuracy are picked out for comparison with IFKNMD-CNN. Furthermore, state-of-the-art fault diagnosis methods based on DL architectures are also compared in this subsection. J Cao et al. proposed a neural architecture search network (NAS), which improves computational efficiency through early stopping and inserts a replay buffer (IRB), and the Pareto efficiency reward function is used to optimize the accuracy, named NAS-PERIRB [37]. L Hou et al. proposed an input feature map (IFM) combined with the residual network (ResNet). The IFM method can extract features without preset parameters [38]. D Wang et al. proposed an attention-based multi-dimensional concatenated convolutional neural network (AMDC-CNN). Important features can be highlighted through the attention mechanism. Multi-dimensional concatenated vibration and torque signals can complement fault features to achieve higher classification accuracy [12].

As shown in Table 7, the accuracy of IFKNMD-CNN is 1.52% higher than the method in [34]. Compared with state-of-the-art DL models, IFKNMD-CNN achieves slightly higher accuracy than NAS-PERIRB and achieves similar performance with IFM-based ResNet and AMDC-CNN. Unlike the above state-of-the-art DL models, this study focuses on proposing an efficient 2D image dataset, so the proposed model does not use the improved CNN architecture for fault classification. The high-quality image dataset used by the proposed model is based on IFK and NMD to extract the most important parts of the signal, and GAF produces high-quality images that preserve absolute temporal relationships. In conclusion, the above proves that IFKNMD-CNN is effective on the Paderborn dataset.

Table 7. Accuracy of existing methods.

Method		*Accuracy* (%)
Machine learning method in [34]	CART	98.3
	RF	98.3
	Ensemble	98.3
NAS-PERIRB [37]		99.43
IFM-based ResNet [38]		99.7
AMDC-CNN [12]		99.8
The proposed model		**99.82**

5. Discussion

Based on the analysis results of Sections 3 and 4, the main advantages of this study are as follows. (1) The IFK can select a narrow frequency band that mainly concentrates fault information and can retain the necessary signal part, thus greatly improving the computational efficiency. As shown in Table 1, the time for NMD to complete the mono-component decomposition significantly reduced from 8691 s to 12 s. Moreover, the IFK has more obvious advantages in analyzing the signals of fault types (KA and KI). As shown in Table 4, the IFK selects a decomposition level of at least 3.5 or above, resulting in a reduction in the signal length of more than 91%. The classification results in Section 4 demonstrate that IFKNMD-CNN achieves high-accuracy fault diagnosis despite imbalanced healthy and faulty samples. (2) The advantage of IFKNMD-CNN's anti-noise capability is based on the IFK and the surrogate data test in NMD. Interfering components such as harmonics and environmental noise in the original signal are first filtered out by the IFK. Then, the carefully designed FT surrogate data and the time–shift surrogate data test the significance and consistency of the signal, respectively. Surrogates that fail the test are rejected, so the noise interference is reduced further.

In addition to the above advantages, IFKNMD-CNN has its limitations. IFK selects the best frequency band by calculating statistical features, and its selection process depends on professional knowledge and the quality of extracted features. Therefore, the performance of the model applied to other datasets may not be as expected.

6. Conclusions

In this paper, a novel model is proposed for intelligent bearing fault diagnosis in rotating machinery. The main contribution of this model is to construct an effective image dataset through the combination of IFK-based NMD and GAF. The proposed model uses the IFK to achieve high computational efficiency and improve SNR. A physically meaningful component of the signal is extracted by NMD. Next, the GAF provides images that preserve the absolute temporal relationship of the signal for CNN to perform fault classification. The Paderborn bearing dataset is used to validate the effectiveness of the model. The validation results show that the proposed model achieves more accurate results as well as robustness to noise when compared with LMD methods based on wavelet denoising and PF selection. IFKNMD-CNN achieves competitive accuracy compared to three state-of-the-art DL methods using the same dataset. Furthermore, the proposed model also demonstrates the generalization ability under different operating conditions. Therefore, IFKNMD-CNN has high potential to help rotating machinery apply intelligent diagnosis under strong noise and different operating conditions. However, this study has not been evaluated under variable speed conditions, and the issue of speed domain adaptability remains to be resolved. Transfer learning has strong potential to learn domain-invariant features under variable speed conditions. Therefore, research on transfer learning is an important direction in the future.

Author Contributions: Methodology, C.-Y.L. and G.-L.Z.; visualization, C.-Y.L. and G.-L.Z.; software, C.-Y.L.; data curation, G.-L.Z.; writing—original draft preparation, G.-L.Z.; validation, G.-L.Z. and T.-A.L.; writing—review and editing, C.-Y.L. and T.-A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gao, S.; Pei, Z.; Zhang, Y.; Li, T. Bearing fault diagnosis based on adaptive convolutional neural network with Nesterov momentum. *IEEE Sens. J.* **2021**, *21*, 9268–9276. [CrossRef]
2. Albrecht, P.F.; Appiarius, J.C.; McCoy, R.M.; Owen, E.L.; Sharma, D.K. Assessment of the reliability of motors in utility applications—Updated. *IEEE Trans. Energy Convers.* **1986**, *EC-1*, 39–46. [CrossRef]
3. Henao, H.; Capolino, G.A.; Fernandez-Cabanas, M.; Filippetti, F.; Bruzzese, C.; Strangas, E.; Pusca, R.; Estima, J.; Riera-Guasp, M.; Hedayati-Kia, S. Trends in fault diagnosis for electrical machines: A review of diagnostic techniques. *IEEE Ind. Electron. Mag.* **2014**, *8*, 31–42. [CrossRef]
4. Guo, L.; Li, N.; Jia, F.; Lei, Y.; Lin, J. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* **2017**, *240*, 98–109. [CrossRef]
5. Dai, X.; Gao, Z. From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Trans. Ind. Informat.* **2013**, *9*, 2226–2238. [CrossRef]
6. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
7. Song, L.; Wang, H.; Chen, P. Vibration-based intelligent fault diagnosis for roller bearings in low-speed rotating machinery. *IEEE Trans. Instrum. Meas.* **2018**, *67*, 1887–1899. [CrossRef]
8. Van, M.; Kang, H.J. Bearing-fault diagnosis using non-local means algorithm and empirical mode decomposition-based feature extraction and two-stage feature selection. *Sci. Meas. Technol.* **2015**, *9*, 671–680. [CrossRef]
9. Wang, J.; Liang, Y.; Zheng, Y.; Gao, R.X.; Zhang, F. An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. *Renew. Energy* **2020**, *145*, 642–650. [CrossRef]
10. Tang, H.; Liao, Z.; Chen, P.; Zuo, D.; Yi, S. A novel convolutional neural network for low-speed structural fault diagnosis under different operating condition and its understanding via visualization. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3501611. [CrossRef]
11. Xu, G.; Liu, M.; Jiang, Z.; Shen, W.; Huang, C. Online fault diagnosis method based on transfer convolutional neural networks. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 509–520. [CrossRef]
12. Wang, D.; Li, Y.; Jia, L.; Song, Y.; Liu, Y. Novel three-stage feature fusion method of multimodal data for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3514710. [CrossRef]
13. Xia, M.; Li, T.; Xu, L.; Liu, L.; De Silva, C.W. Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks. *IEEE/ASME Trans. Mechatron.* **2018**, *23*, 101–110. [CrossRef]
14. Zhao, R.; Yan, R.Q.; Chen, Z.H.; Mao, K.Z.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
15. Zhang, W.; Li, C.; Peng, G.; Chen, Y.; Zhang, Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mech. Syst. Signal Process.* **2018**, *100*, 439–453. [CrossRef]
16. Jia, F.; Lei, Y.; Lu, N.; Xing, S. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mech. Syst. Signal Process.* **2018**, *110*, 349–367. [CrossRef]
17. Huang, W.; Cheng, J.; Yang, Y.; Guo, G. An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis. *Neurocomputing* **2019**, *359*, 77–92. [CrossRef]
18. Li, X.; Zhang, W.; Ding, Q. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Rel. Eng. Syst. Saf.* **2019**, *182*, 208–218. [CrossRef]
19. Darong, H.; Lanyan, K.; Bo, M.; Ling, Z.; Guoxi, S. A new incipient fault diagnosis method combining improved RLS and LMD algorithm for rolling bearings with strong background noise. *IEEE Access* **2018**, *6*, 26001–26010. [CrossRef]
20. Zhao, D.; Zhang, H.; Liu, S.; Wei, Y.; Xiao, S. Deep rational attention network with threshold strategy embedded for mechanical fault diagnosis. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3519715. [CrossRef]
21. Tang, X.; Xu, Z.; Wang, Z. A novel fault diagnosis method of rolling bearing based on integrated vision transformer model. *Sensors* **2022**, *22*, 3878. [CrossRef] [PubMed]
22. Qiao, H.; Wang, T.; Wang, P.; Zhang, L.; Xu, M. An adaptive weighted multiscale convolutional neural network for rotating machinery fault diagnosis under variable operating conditions. *IEEE Access* **2019**, *7*, 118954–118964. [CrossRef]

23. Qin, C.; Jin, Y.; Tao, J.; Xiao, D.; Yu, H.; Liu, C.; Lei, J.; Liu, C. DTCNNMI: A deep twin convolutional neural networks with multi-domain inputs for strongly noisy diesel engine misfire detection. *Meas. J. Int. Meas. Confed.* **2021**, *180*, 109548. [CrossRef]
24. Qiao, M.; Yan, S.; Tang, X.; Xu, C. Deep convolutional and LSTM recurrent neural networks for rolling bearing fault diagnosis under strong noises and variable loads. *IEEE Access* **2020**, *8*, 66257–66269. [CrossRef]
25. Jin, G.; Zhu, T.; Akram, M.W.; Jin, Y.; Zhu, C. An adaptive anti-noise neural network for bearing fault diagnosis under noise and varying load conditions. *IEEE Access* **2020**, *8*, 74793–74807. [CrossRef]
26. Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors* **2017**, *17*, 425. [CrossRef]
27. Iatsenko, D.; McClintock, P.V.E.; Stefanovska, A. Nonlinear mode decomposition: A noise-robust, adaptive decomposition method. *Phys. Rev. E* **2015**, *92*, 032916. [CrossRef]
28. Iatsenko, D.; McClintock, P.V.E.; Stefanovska, A. Linear and synchrosqueezed time-frequency representations revisited. part ii: Resolution reconstruction and concentration. *Digit. Signal Process.* **2015**, *42*, 1–26. [CrossRef]
29. Schreiber, T.; Schmitz, A. Surrogate time series. *Physica D* **1999**, *142*, 346–382. [CrossRef]
30. Sheppard, L.W.; Stefanovska, A.; McClintock, P.V.E. Detecting the harmonics of oscillations with time-variable frequencies. *Phys. Rev. E* **2011**, *83*, 016206. [CrossRef]
31. Wang, Y.; Tang, B.; Qin, Y.; Huang, T. Rolling bearing fault detection of civil aircraft engine based on adaptive estimation of instantaneous angular speed. *IEEE Trans. Ind. Informat.* **2020**, *16*, 4938–4948. [CrossRef]
32. Feng, Z.; Zhang, D.; Zuo, M.J. Adaptive mode decomposition methods and their applications in signal analysis for machinery fault diagnosis: A review with examples. *IEEE Access* **2017**, *5*, 24301–24331. [CrossRef]
33. Wang, Z.; Oates, T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In Proceedings of the Workshops at AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 40–46.
34. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In Proceedings of the PHM Society European Conference, Bilbao, Spain, 5–8 July 2016; pp. 5–8.
35. Antoni, J. Fast computation of the Kurtogram for the detection of transient faults. *Mech. Syst. Signal Process.* **2007**, *21*, 108–124. [CrossRef]
36. Lee, C.-Y.; Zhuo, G.-L. Effective rotor fault diagnosis model using multilayer signal analysis and hybrid genetic binary chicken swarm optimization. *Symmetry* **2021**, *13*, 487. [CrossRef]
37. Cao, J.; Ma, J.; Huang, D.; Yu, P. Finding the optimal multilayer network structure through reinforcement learning in fault diagnosis. *Measurement* **2022**, *188*, 110377. [CrossRef]
38. Hou, L.; Jiang, R.; Tan, Y.; Zhang, J. Input feature mappings-based deep residual networks for fault diagnosis of rolling element bearing with complicated dataset. *IEEE Access* **2020**, *8*, 180967–180976. [CrossRef]

sensors

MDPI

Article

Implementation of Aging Mechanism Analysis and Prediction for XILINX 7-Series FPGAs with a 28-nm Process

Zeyu Li, Zhao Huang *, Quan Wang, Junjie Wang and Nan Luo

School of Computer Science and Technology, Xidian University, Xi'an 710071, China; zeyuli@stu.xidian.edu.cn (Z.L.); qwang@xidian.edu.cn (Q.W.); junjiewang@stu.xidian.edu.cn (J.W.); nluo@xidian.edu.cn (N.L.)

* Correspondence: z_huang@xidian.edu.cn; Tel.: +86-1879-261-0378

Abstract: Commercial off-the-shelf (COTS) field-programmable gate arrays (FPGAs) with a 28-nm process have become popular devices for computing systems. Although current generation FPGAs have advantages over previous models, the phenomenon of circuit aging has become more significant with the sharp reduction in the process size of FPGAs. Aging results in FPGA performance degradation over time and, ultimately, hard faults. However, few studies have focused on understanding aging mechanisms or estimating the aging trend of 28-nm FPGAs. For this, we used a ring oscillator (RO)-based test structure to extract data and build a dataset that could be used to predict aging trends and determine the primary aging mechanisms of 28-nm FPGAs. Moreover, we proposed a correction method to correct temperature-induced measurement errors in accelerated tests. Furthermore, we employed four machine learning (ML) technologies that were based on accurate measurement datasets to predict FPGA aging trends. In the experiment, 24 XILINX 7-series FPGAs (28 nm) were evaluated for 10+ years of circuit operation using accelerated tests. The results showed that the aging effects of negative-bias temperature instability (NBTI) was the primary aging mechanism. The correction method proposed in this paper could effectively eliminate measurement errors. In addition, the minimum prediction error rate of the ML model was only 0.292%.

Keywords: FPGA; aging mechanism; NBTI; measurement error correction; aging prediction; machine learning

Citation: Li, Z.; Huang, Z.; Wang, Q.; Wang, J.; Luo, N. Implementation of Aging Mechanism Analysis and Prediction for XILINX 7-Series FPGAs with a 28-nm Process. *Sensors* **2022**, *22*, 4439. https://doi.org/10.3390/s22124439

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 13 May 2022
Accepted: 8 June 2022
Published: 12 June 2022

1. Introduction

Commercial off-the-shelf (COTS) field-programmable gate arrays (FPGAs) with a 28-nm process have become popular devices for computing systems. Although current generation FPGAs have advantages over previous models, the continuous scaling of devices to deep nanotechnology and the inexorable reduction in supply voltage significantly challenge the reliability assurance that is related to device aging [1–3]. Aging results in FPGA performance degradation over time and, ultimately, hard faults. Hence, it is essential to understand the main aging mechanisms of FPGAs [4–6]. Meanwhile, estimating the aging trends of age-related faults before they occur is crucial for developing aging prevention/mitigation actions to avoid circuit failures [7,8].

To effectively solve the above problems, many efforts have been devoted to aging tests for the analysis of aging mechanisms and the prediction of aging trends of FPGAs. The increase in path delay is the primary indicator of FPGA aging degradation. Hence, measuring the variations in path delay can quantify the aging degree of a circuit. For a long time, actual on-chip measurements and sensor-based aging monitoring have been the mainstream methods [2,9–14]. Almost all of these methods employ ring oscillator (RO)-based circuits to measure path delay. However, the test processes of the above methods easily affect the measured delay results and cause errors. Therefore, it is essential to correct measurement errors to obtain accurate data. In terms of aging prediction, most studies have used physical aging models to simulate the aging degradation of transistors

or look-up tables (LUTs) [15–17]. However, the parameters of such models are difficult to determine. Furthermore, some studies have predicted the aging of circuits based on machine learning (ML) [18,19]. Nevertheless, these methods only focus on predicting the path delay degradation that is related to bias temperature instability (BTI).

To make up for the limitations of previous research, we performed an on-chip, accelerated aging test to obtain the aging-related data of 28-nm FPGAs. Meanwhile, we also improved a measurement method to correct measurement errors that were caused by the accelerated experiment. Based on the above work, on the one hand, we investigated the primary aging mechanisms of 28-nm FPGAs; on the other hand, we employed a variety of ML technologies to predict the aging trends of FPGAs. In summary, we achieved the following novel contributions:

- We performed an on-chip, accelerated aging test to observe the effects of different stress signals and LUT configurations on FPGA aging, which showed how the frequencies of ROs change with aging and which aging mechanisms mainly affect 28-nm FPGAs;
- A measurement method was improved to correct measurement errors that were caused by the accelerated experiment and the corrected data were used for the analysis of the aging effects and the training of the aging prediction model;
- A variety of machine learning technologies were employed to predict the aging trends of FPGAs to evaluate the effectiveness of the ML models for the prediction of FPGA aging trends.
- The experimental results, based on a group of 28-nm XILINX 7-series FPGAs, showed that negative BTI (NBTI) was the main aging mechanism; moreover, the correction method proposed in this paper could effectively rectify measurement errors and in terms of aging prediction, the XGBoost-based ML model was competent for fitting the actual aging trends of FPGAs.

The structure of this paper is organized as follows. In Section 2, we review the important aging mechanisms of ICs and describe related works on aging tests and the aging prediction of FPGAs. Section 3 presents the aging test implementation using FPGAs and proposes the error correction method. The experimental results are presented in Section 4, followed by the conclusion in Section 5.

2. Background and Related Work

2.1. Aging Mechanisms

Circuit aging refers to the degradation of some of the characteristic hardware parameters in integrated circuits (ICs) over time. It can be summarized as the increase in threshold voltage that is caused by transistor aging, which eventually leads to transistor failure, and the increase in resistance that is caused by metal wire aging, which eventually leads to fracture. The aging mechanisms of transistors and interconnects are dominated by four main effects at the nanoscale: bias temperature instability (BTI), hot carrier injection (HCI), time-dependent dielectric breakdown (TDDB), and electromigration (EM).

BTI is considered to be the main limiting factor of the lifetime of nanoscale complementary metal oxide semiconductor (CMOS) devices and is divided into positive and negative BTI (PBTI/NBTI) [20,21]. HCI is due to the strong channel electric fields near the drain in the channel, which causes the carriers to cross the Si–SiO_2 barrier and inject into the oxide medium to form traps and results in the degradation of the threshold voltage [22,23]. TDDB causes local tunnel breakdown and eventually causes dielectric breakdown, which usually leads to catastrophic hard failure. EM is a mechanism that affects the interconnects and induces open circuits (due to voids) or short circuits (due to hillocks).

2.2. Aging Tests on FPGAs

FPGA aging degradation can manifest itself as an increase in the probability of transient/permanent failures [24] or as a change in timing. One of the most intuitive and easily observable indications is the increase in the path delay of the circuit. Hence, measuring the variations in path delay can quantify the aging degree of a circuit. In the early stages, a

transition probability-based delay measurement is the primary method that is used [25]. However, the delay data obtained by this method are not accurate enough since they usually evaluate the worst-case path delay.

With the popularization of built-in self-tests (BISTs) in IC tests, actual on-chip measurements and sensor-based aging monitoring have become the mainstream methods [2,9–14,26,27]. Naouss et al. [2] established a low-cost test platform to evaluate FPGA reliability, which supports aging delay measurements for multiple FPGAs at the same time. Miyake et al. and Xiang et al. [9,10] proposed a measurement method based on ROs concerning on-chip delay, which is suitable for field testing. Refs. [11–14] employed aging sensors to monitor the delays in critical circuit paths to evaluate FPGA aging. Almost all of these methods can obtain relatively accurate delay data and their measurements are based on RO circuits. Hence, this study also employed RO-based measurement circuits to test FPGA aging.

2.3. Aging Prediction of FPGAs

Most early studies used physical aging models to predict the aging degradation of transistors and LUTs. Morales et al. [15] developed a general simulation environment to implement FPGA circuits that can predict the LUT propagation delay of digital circuits. Jang et al. [16] proposed an on-chip aging sensor circuit to predict and detect circuit failures caused by the effects of BTI and HCI aging on digital circuits. Yu et al. [17] proposed a fast time-zero aging prediction and predictive screening methodology based on a novel on-chip architecture, named ZeroScreen. However, the implementation of the above methods usually depends on the transistor or LUT model. Therefore, it is difficult to determine the appropriate formula parameters.

To date, some studies have predicted the aging of circuits based on ML [18,19]. For example, Karimi et al. [18] proposed a general-purpose IC aging prognosis approach that considers a comprehensive set of IC operating conditions, including workload, usage time, and operating temperature. Vijayan et al. [19] proposed a method to perform low-cost and fine-grained workload-induced stress monitoring for accurate age-induced delay prediction. However, these methods only focus on predicting the degradation of BTI-related path delay. In addition, they also have to depend on logic simulation to obtain characteristic values and labels as inputs for the prediction model. In contrast, our method directly exploits the measured data to train the ML-based aging prediction model. As a result, we could predict FPGA aging without depending on physical aging models.

3. Aging Test Implementation for FPGAs

3.1. Design of Test Solution

In this study, the on-chip aging test was performed using an RO circuit. Due to the self-oscillation characteristics of the RO, the change in its frequency could characterize the aging degradation of FPGAs.

Figure 1 shows the RO-based structure. The test circuit had two working modes: accelerated aging mode (0) and test mode (1). The user sent the status-control bit signal to the circuit through the UART when switching modes. When mode = 1, the circuit was in an open-loop state to accelerate its aging under test conditions by inputting a signal of a specific waveform as a stress signal. This could be a static signal (DC0, DC1) or a signal that was generated via a PLL of the FPGA, for which the user determined the frequency and duty cycle. When mode = 0, the aging state was be measured. As the circuit was in a closed-loop condition at this point, a measurement method based on the ring oscillator was employed and the counter produced the corresponding frequency. During the test, we also used the XADC IP core [28] to periodically monitor the core temperature and analyze the influence of temperature on measurement errors.

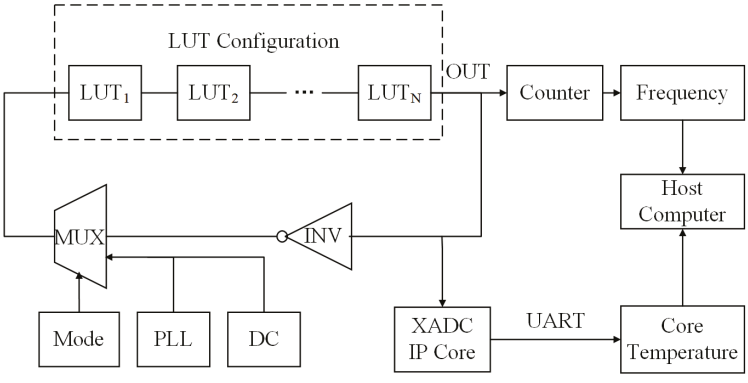


Figure 1. RO-based test structure.

The logic function of the test is shown in Figure 2. The core of the test was the controller module, which was responsible for coordinating the whole test process. The core voltage supply module provided the required working voltage for the FPGA. The RS232 module was responsible for the communication between the FPGA and the PC. The input was the configuration file and stress signal of the circuit under test (CUT) and the output was the frequency value of the CUT.

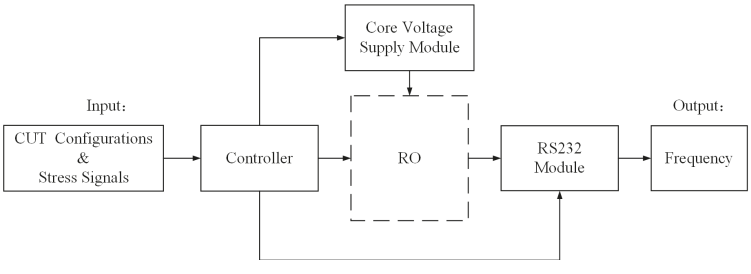


Figure 2. Logic function module of the aging test.

3.2. Accelerated Aging Conditions

An accelerated test refers to the accelerated degradation of a tested product by strengthening the test conditions, under the premise of ensuring that the failure mechanisms of the product are not changed, to obtain the necessary information in a relatively short period of time [29]. The aging speed of FPGAs is normally limited and long-term aging tests cannot be achieved within set time parameters. Hence, it is incredibly important to carry out accelerated tests [30]. In line with the principle regarding the aging mechanisms of BTI, HCI, and the theoretical acceleration model, the aging speed was directly related to the working voltage and temperature of the circuit and their relationships could be expressed as in the following formulae [31]:

$$t_f \propto V_{gs}^{-\gamma} \tag{1}$$

$$t_f \propto exp(\frac{E_a}{kT}) \tag{2}$$

where t_f represents the estimated duration for which the circuit can operate reliably. Based on the evidence from the available work in the literature [31], the value of γ is usually 6–8, E_a is approximately 1.5 eV, and k is Boltzmann's constant. T stands for the operating temperature of the circuit, while V_{gs} is the gate-source voltage of the transistor.

The objects of this accelerated aging test were XILINX 7-series FPGAs. According to the FPGA manual, the range of supply voltage, without causing damage, is 0.5 V to 1.1 V, the general working voltage is 1.0 V, the working temperature is between −40 °C and +100 °C, and the general working environment temperature is 27 °C. Based on aging theory and the test conditions, Table 1 presents the theoretical power supply voltages and operating temperatures for the aging tests and the estimates of the acceleration under these conditions.

Table 1. Conditions of accelerated tests.

Factor	Relationship	Stress Condition	Acceleration
Core Voltage Supply	$t_f \propto V_{gs}^{-\gamma}$	1.1 V	≈10×
Temperature	$t_f \propto exp(\frac{E_a}{kT})$	373 K	≈2×
Voltage and Temperature			≈20×

3.3. Correction Method for Measurement Errors

The aging degree of a device is aggravated by increases in operation time, which indicates the cumulative growth of the circuit path delay. The acceleration experiment was carried out in a high-temperature environment and the delay that was measured on-chip was affected by the temperature. Assuming that the initial delay of the circuit was D_0, the measured value was the sum of the initial delay, the aging delay D_{aging} of the circuit, and the temperature-related error value D_{error}:

$$D_{measure} = D_0 + D_{aging} + D_{error} \tag{3}$$

At this time, the measurement value could not reflect the real delay of the circuit. Thus, the influence of delay variation due to temperature change had to be eliminated to obtain temperature-independent delay measurements. In this regard, we researched the error correction method. It was assumed that the time delay caused by aging would not increase over a concise period, i.e., $\lim_{\Delta t \to 0} \Delta D_{aging} = 0$, where ΔD_{aging} is the delay variation caused by aging. Therefore, the delay variation from sampling at different temperatures was the delay error caused by temperature, i.e., $\Delta D = D_{error} = D_{measure} - D_0$. At this time, the correlation coefficient $\lambda = \Delta D / \Delta T$ was introduced, which represented the relationship between the change in measurement delay and the change in temperature. When λ was a constant value (i.e., the variation in delay error caused by temperature was in a fixed proportion to the variation in temperature), the measurement error could be corrected by the λ value, correction value $D_{correct} = \lambda \Delta T$, and real aging delay $D_{aging} = D_{measure} \pm D_{correct}$. Then, the research focused on computing the value of λ.

To obtain λ, we first measured the initial circuit delay d_0 and then constantly changed the core temperature (CT) and synchronously measured the change in the on-chip circuit delay. Meanwhile, we recorded the D value between the current and initial temperatures. To reduce the difference that was caused by this process, the experiment was carried out on six FPGAs, with each FPGA running the same CUT. We simultaneously set the temperature, recorded the relevant data, and obtained the average value of λ across the six groups of data. Before that, we also determined the relationship between the external environment temperature (AT) and the CT in order to accurately adjust the core temperature. Figure 3 shows the relationship between the ambient temperature and the core temperature throughout the experimental measurements.

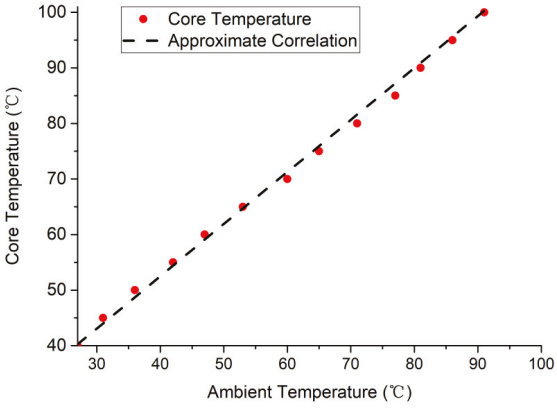


Figure 3. The relationship between the ambient temperature and the core temperature.

The red circle in Figure 3 represents the measured CT and it can be seen that the CT had a linear relationship with ΔT. It was found that they accorded with $CT = 0.9375 \times AT + 15$ (dotted line) by calculation. It should be noted that the environment was a sealed aging test chamber. When the AT reached 91 °C, the CT could reach 100 °C, which was the upper limit of the CT of the FPGAs. We could adjust the CT accurately, according to the formula. In the experiment, the initial temperature was 27 °C. We adjusted the core temperature from 40 °C to 100 °C in 5-°C intervals and calculated the corresponding λ temperature, which was denoted as λ_1~λ_{12}. The results are shown in Table 2.

Table 2. The results of λ.

Coefficient	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Value	0.0087	0.0083	0.0069	0.0074	0.0063	0.0063
Coefficient	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}
Value	0.0053	0.0038	0.0033	0.0027	0.0016	0.0001

It can be seen that the λ values that corresponded to different temperatures obviously changed, which indicated that the correction method mentioned above could not be used directly. We proposed the hypothesis that when the λ value does not change with the increase in duration at the same temperature, the delay variation measured over two time periods is equal to the delay variation that is caused by aging. Then, the actual aging delay could be corrected by calculating the correction value $D_{correct} = \lambda_T \times \Delta T$, where λ_T is the coefficient at the current temperature. To verify the feasibility of this method, we calculated the λ coefficient at different aging times. We designed a 1000-h experiment and calculated the λ coefficients every 100 h. The λ coefficients corresponded to 60 °C (λ_1), 80 °C (λ_2), and 100 °C (λ_3). Similarly, to reduce the difference caused by the process, the experiment was carried out on 10 FPGAs, with each FPGA running the same CUT. We simultaneously set the temperature, recorded the relevant data, and obtained the average value across the ten data groups. The experimental results are shown in Table 3.

Table 3. The results of λ over time.

Coefficient	100 h	200 h	300 h	400 h	500 h
λ_1	0.0074	0.0076	0.0075	0.0074	0.0075
λ_2	0.0038	0.0039	0.0038	0.0039	0.0038
λ_3	0.0000	0.0002	0.0000	0.0001	0.0000
Coefficient	**600 h**	**700 h**	**800 h**	**900 h**	**1000 h**
λ_1	0.0075	0.0074	0.0074	0.0074	0.0074
λ_2	0.0038	0.0040	0.0038	0.0039	0.0038
λ_3	0.0000	0.0000	0.0001	0.0000	0.0000

It can be seen from the table that there were slight differences (measurement errors) in the measured λ values at different temperatures. Generally speaking, it could be proved that the temperature-dependent delay did not change with the duration increase. In practice, we calculated the corresponding λ value at the temperature that corresponded to the core temperature and then, we could correct any errors in the measurement. By restoring the measurement errors, the accurate aging delay could be obtained by way of on-chip measurements. The chip could be continuously accelerated without waiting for the temperature to return to the average temperature to obtain accurate measurements. It is evident that each heating and cooling process was time-consuming and that the critical data were unstable.

4. Test Results and Analysis

4.1. Experimental Setup

We used 24 XILINX 7-series FPGAs (28 nm) for the aging tests in our experiment. The host computer was a Xeon(R) Silver 4116 (2.10 GHz) CPU with 32 GB DDR4 RAM, which was running Windows 10. The reconfiguration fabrics of each FPGA were divided into 16 reconfigurable regions to execute the CUTs.

To understand the primary aging mechanisms of 28-nm FPGAs, we combined different stress signals and LUT configurations as the test conditions. In this experiment, five common frequencies (DC0, DC1, 50 MHz, 100 MHz, and 300 MHz) and three duty cycles (DC25, DC50, and DC75) were selected as the combined stress signals and were input into the CUTs. The LUTs of the ROs were configured as BUFFER, AND, XOR, and INV and were executed in each of the four groups of chips. Hence, degradation due to certain test conditions was the mean value of the degradation of the six circuits under test.

The conditions of V and T under this setting were approximately equal in order to eliminate any differences in the manufacturing process. The value of each data point was the average value of the same six CUTs. Moreover, the voltage was provided by external stabilized power and the high temperatures were produced by a 101-0B high-temperature test chamber, as shown in Figure 4, which was capable of providing a stable temperature environment for the test from 50 °C to 300 °C, thus meeting the needs of the accelerated degradation tests.

To evaluate the ML-based models for the application of FPGA aging prediction, we employed four ML technologies (XGBoost, SVM, LR, and ANN) to model the reconfiguration circuits. In the experiments, we extracted the data from all 24 XILINX 7-series FPGAs from the aging tests and aging simulation experiments to build our dataset (frequency, duty cycle, operation time, LUT configuration, delay variation, etc.) and this dataset was then used to train and test the prediction model. We used the root mean squared error (RMSE) as the evaluation metric in this experiment.

Figure 4. Operation state of the high-temperature test chamber.

4.2. Influence of Stress Signals on FPGA Aging

Here, we present the influences of different stress signals on FPGA aging degradation and their analysis to find the primary aging mechanisms of 28-nm FPGAs.

4.2.1. The Influence of Frequency

Dynamic stresses (50 Mhz, 100 Mhz, and 300 Mhz) and static stresses (DC0 and DC1) that related to different operating frequencies were selected as the inputs for the stress signals. Figure 5 shows the frequency degradation of the ROs under other test conditions. As expected, the degradation that was caused by the NBTI and HCI mechanisms increased as the temperature rose. After 1000 h of accelerated testing, we found that the degradation was 1.8% at a working temperature of 100 °C and 0.9% at a working temperature of 25 °C. However, we did not observe regularity in the aging degradation that was caused by dynamic AC stress. One possible explanation could be that this degradation results from the combined effect of two aging mechanisms: NBTI and HCI. Existing studies have demonstrated that the aging effects of NBTI decrease with increasing stress frequency, while the aging effects of HCI increase with increasing stress frequency [32,33]. When the stress frequency changes, these two aging mechanisms change in opposite directions at the same time. Therefore, it could not be analyzed whether there was a (positive or negative) correlation between the change in frequency and the aging degradation.

4.2.2. The Influence of Duty Cycle

Three duty cycles (25%, 50%, and 75%) were selected as the stress signals input. Figure 6 shows the frequency degradation of the ROs due to different AC stress signals with the different duty cycles. For the same stress signal frequency, the 25% duty cycle had a more significant drop than the 50% duty cycle, while the 50% duty cycle had a more significant drop than the 75% duty cycle. We could see that this difference was more pronounced at higher temperatures (100 °C vs. 25 °C) and at higher stress signal frequencies (300 MHz vs. 10 MHz). This could be explained by the fact that the period of the low-frequency signal was long enough to restore the NBTI aging mechanism to some extent.

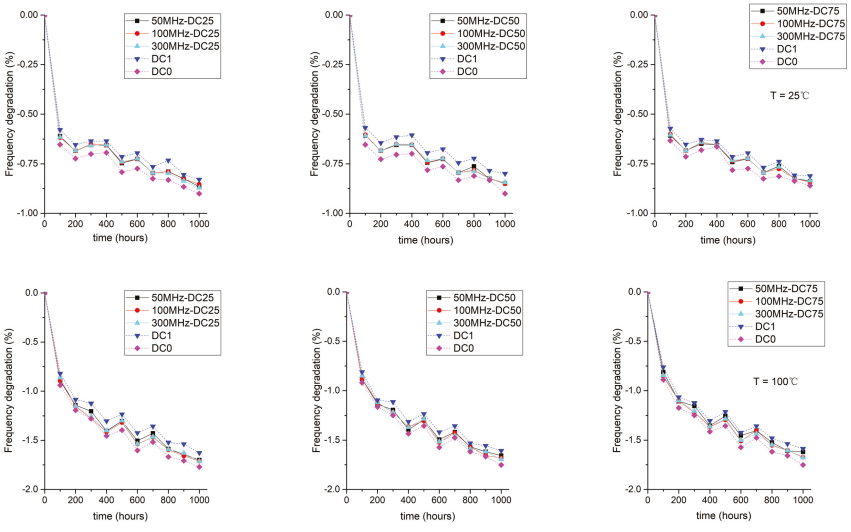


Figure 5. Impacts of stress signal frequency on the frequency degradation of ROs.

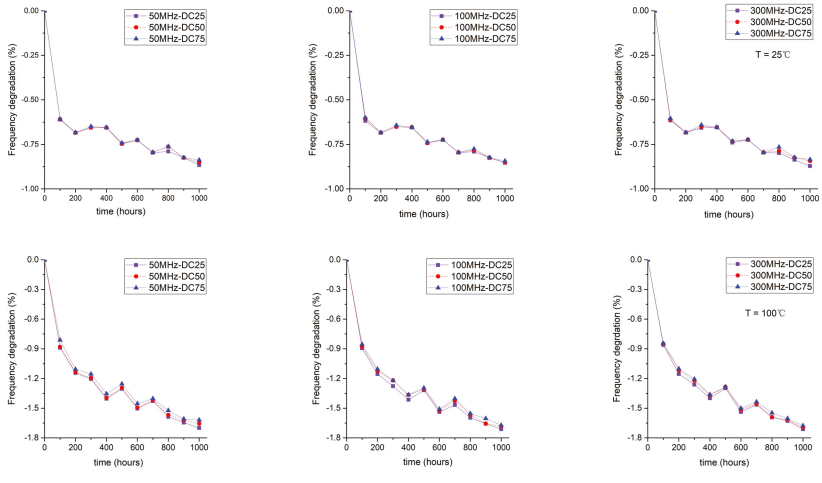


Figure 6. Impacts of duty cycle on the frequency degradation of ROs.

4.3. Evaluation of Correction Method

In the experiments, we set the core temperature to 90 °C. We sampled 10 data groups and measured the path delay at 100 h, 200 h, 300 h, 400 h, 500 h, 600 h, 700 h, 800 h, 900 h, and 1000 h. To eliminate measured errors/noise points, the experiment was carried out synchronously on 10 FPGAs and each data point was the mean value across the 10 groups of measured data. After the measurements, we corrected the errors and recorded the corrected data. To evaluate the effectiveness of the correction method, we used ModelSim to simulate the aging of the XILINX 7-series FPGAs and recorded the delay data that corresponded to the simulation time points. As shown in Figure 7, it was found that the difference between the corrected delay and the simulation delay was within 1%, which proved that the correction method was effective.

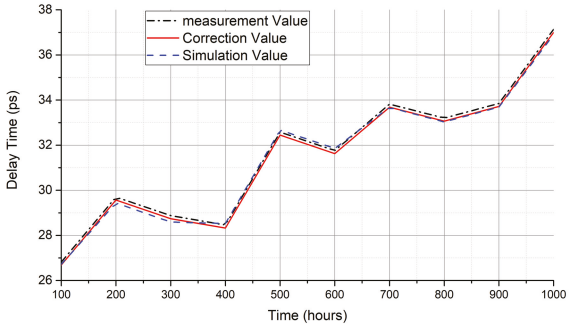


Figure 7. Comparison of the correction delay and simulation delay.

4.4. Results of Aging Prediction

The results of the RMSE of the XGBoost, SVM, LR, and ANN models are presented in Figure 8. It can be seen that the RMSE of the ANN was very stable, but there were still significant errors when the predicted values of frequency degradation were low. The RMSE of LR and SVM were relatively high and there were also significant prediction errors. Compared to the other three models, the RMSE of XGBoost was minimal. The increase in RMSE was due to the predicted frequency degradation value also increasing, but the prediction error did not change significantly. The mean error rate of the XGBoost prediction was only 0.292%.

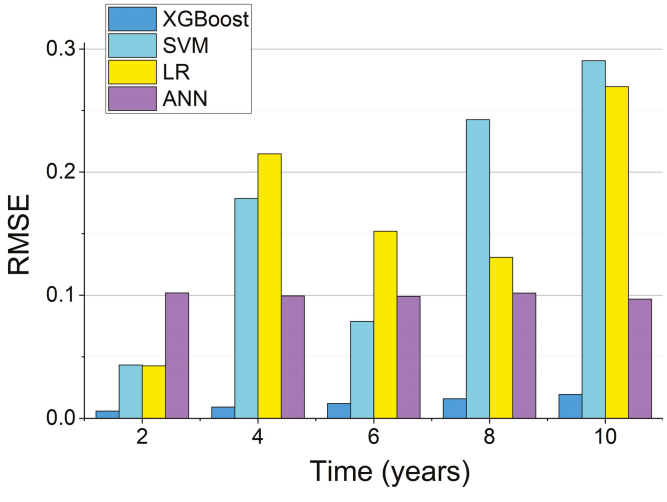


Figure 8. RMSEs of the different ML models.

Figure 9 presents the aging prediction results of the four ML models under different stress signals and LUT configurations of CUTs. The base represents the measured aging degradation. As the results show, the aging trends that were predicted by all ML models were similar to the actual aging trends (red), particularly the prediction of the XGBoost model, which almost entirely coincided with the actual aging trend. Hence, the above experiments illustrated that it would be very feasible to use the ML models to predict the aging degradation of FPGAs.

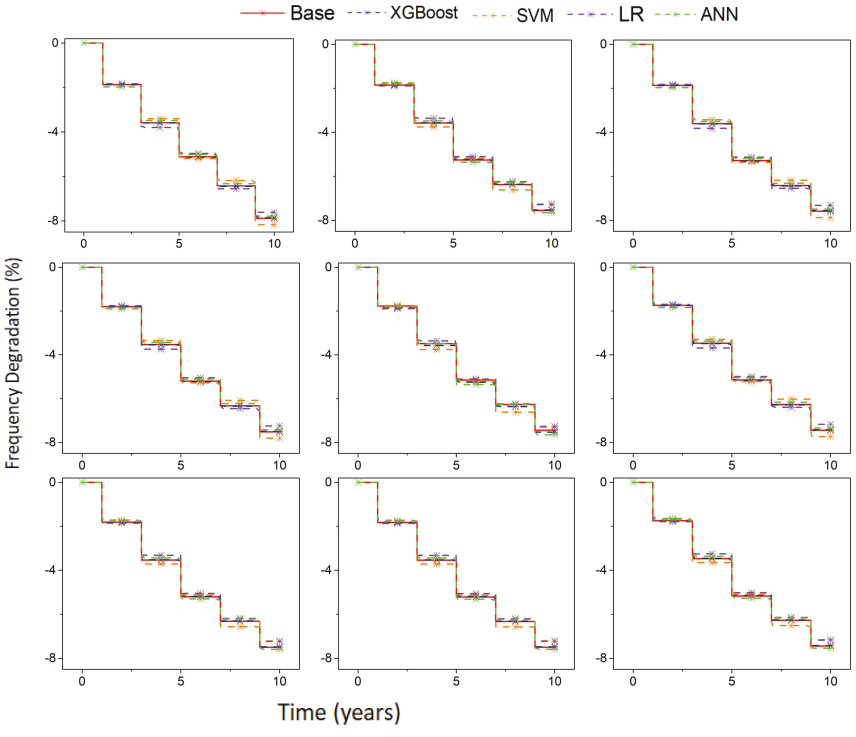


Figure 9. Aging prediction results of the different ML models under different stress signals and LUT configurations.

4.5. Discussion

With the shrinking CMOS manufacturing process, NBTI has proven to be the most important aging mechanism. While existing studies have validated this conclusion by performing aging tests on XILINX ARTIX7 FPGAs, we attempted to further validate this conclusion by performing aging tests on a larger number/type of XILINX 7-series FPGAs. The experiments in this paper also demonstrated that NBTI is the most important aging mechanism for 28-nm FPGAs. Moreover, it is also worth noting that there were two other contributions of this paper for the community: the error correction method for the aging test and the prediction of FPGA aging based on machine learning models.

To reduce the duration of FPGA aging tests, common practice is to place the device in a high-temperature test box to accelerate aging. However, due to the measurement errors that are caused by high-temperature environments, the delay that is measured does not reflect the actual degree of aging degradation of the device. To this end, we proposed an error correction method for the aging tests. Our experiments showed that the error correction method proposed in this paper is effective. In addition, as far as the literature that was reviewed by the authors is concerned, there are few studies on predicting FPGA aging based on ML. To evaluate the aging trends of devices more efficiently, we explored the use of machine learning models to build an FPGA aging prediction model. Through experimental evaluation, the aging prediction model that was based on machine learning can better fit the real aging trends of devices.

5. Conclusions and Future Work

In this work, we studied the main aging mechanisms of 28-nm FPGAs. Different stress signals and LUT configurations were applied in aging tests. The results showed

that NBTI is the main reason for FPGA aging degradation. To collect accurate aging data, we further analyzed the influence of temperature on measurement errors and proposed an error correction method. The results showed that the difference between the corrected measurement results and the simulation results was less than 1%, thereby proving that the correction method is efficient. Moreover, we employed four ML models that were trained using aging data to predict FPGA aging. Among them, the mean error rate of the XGBoost prediction was only 0.292%, which proves that it would be very feasible to use the ML model to predict the aging trends of FPGAs.

In future work, we will evaluate the effectiveness of the error correction and aging prediction methods that were proposed in this paper more comprehensively by testing different types of FPGAs. In addition, we will investigate more age-related features (e.g., failure rate) and incorporate them into the prediction models to further improve the accuracy of the model prediction. For the established aging prediction model, we will apply it to preventive maintenance in order to evaluate and predict the trends and extent of the circuit aging of FPGAs under different stress signals and LUT configurations. This will support the rational use of age-aware scheduling strategies to achieve aging mitigation.

Author Contributions: Conceptualization, Q.W. and Z.H.; methodology, Z.H., N.L. and Z.L.; software, Z.L. and J.W.; validation, Z.H., N.L. and Z.L.; writing—original draft preparation, Z.L. and J.W.; writing—review and editing, Z.H. and J.W.; supervision, Q.W.; project administration, Q.W.; funding acquisition, Q.W., N.L. and Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grant number 61972302, in part by the Fundamental Research Fund for the Central Universities under grant number XJS220306, in part by the Natural Science Basic Research Program of Shaanxi under grant number 2022JQ-680, in part by the Key Research and Development Program of Shannxi Province under grant numbers 2021GY-086 and 2021GY-014, and in part by the Key Laboratory of Smart Human–Computer Interaction and Wearable Technology of Shaanxi Province.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stott, E.A.; Wong, J.S.; Sedcole, P.; Cheung, P.Y. Degradation in FPGAs: Measurement and modelling. In Proceedings of the 18th Annual ACM/SIGDA International Symposium on Field Programmable Gate Arrays, Monterey, CA, USA, 21–23 February 2010; pp. 229–238.
2. Naouss, M.; Marc, F. Design and implementation of a low cost test bench to assess the reliability of FPGA. *Microelectron. Reliab.* **2015**, *55*, 1341–1345. [CrossRef]
3. Kiamehr, S.; Weckx, P.; Tahoori, M.; Kaczer, B.; Kukner, H.; Raghavan, P.; Groeseneken, G.; Catthoor, F. The impact of process variation and stochastic aging in nanoscale VLSI. In Proceedings of the 2016 IEEE International Reliability Physics Symposium (IRPS), Pasadena, CA, USA, 17–21 April 2016; pp. CR–1–1–CR–1–6.
4. Karapetyan, S.; Schlichtmann, U. Integrating aging aware timing analysis into a commercial STA tool. In Proceedings of the VLSI Design, Automation and Test (VLSI-DAT), Hsinchu, Taiwan, 27–29 April 2015; pp. 1–4.
5. Dogan, H.; Forte, D.; Tehranipoor, M.M. Aging analysis for recycled FPGA detection. In Proceedings of the 2014 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), Amsterdam, The Netherlands, 1–3 October 2014; pp. 171–176.
6. Maiti, A.; McDougall, L.; Schaumont, P. The impact of aging on an FPGA-based physical unclonable function. In Proceedings of the 2011 21st International Conference on Field Programmable Logic and Applications, Chania, Greece, 5–7 September 2011; pp. 151–156.
7. Ebrahimi, M.; Sadeghi, R.; Navabi, Z. LUT input reordering to reduce aging impact on FPGA LUTs. *IEEE Trans. Comput.* **2020**, *69*, 1500–1506. [CrossRef]
8. Alam, M.M.; Tehranipoor, M.; Forte, D. Recycled FPGA detection using exhaustive LUT path delay characterization and voltage scaling. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2019**, *27*, 2897–2910. [CrossRef]
9. Miyake, Y.; Sato, Y.; Kajihara, S. On-Chip Delay Measurement for In-Field Test of FPGAs. In Proceedings of the 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), Kyoto, Japan, 1–3 December 2019; pp. 130–1307.
10. Xiang, Z.J.; Liu, W.; Wang, L.h; Wang, L.L. A System for FPGA Aging Test. In Proceedings of the 2018 10th International Conference on Communications, Circuits and Systems (ICCCAS), Chengdu, China, 22–24 December 2018; pp. 471–474.

11. Ebrahimi, M.; Ghaderi, Z.; Bozorgzadeh, E.; Navabi, Z. Path selection and sensor insertion flow for age monitoring in FPGAs. In Proceedings of the 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, Germany, 14–18 March 2016; pp. 792–797.
12. Ghaderi, Z.; Ebrahimi, M.; Navabi, Z.; Bozorgzadeh, E.; Bagherzadeh, N. SENSIBle: A highly scalable sensor design for path-based age monitoring in FPGAs. *IEEE Trans. Comput.* **2016**, *66*, 919–926. [CrossRef]
13. Firouzi, F.; Ye, F.; Chakrabarty, K.; Tahoori, M.B. Aging-and variation-aware delay monitoring using representative critical path selection. *ACM Trans. Des. Autom. Electron. Syst. (TODAES)* **2015**, *20*, 1–23. [CrossRef]
14. Valdes-Pena, M.D.; Freijedo, J.F.; Rodriguez, M.J.M.; Rodriguez-Andina, J.J.; Semião, J.; Teixeira, I.M.C.; Teixeira, J.P.C.; Vargas, F. Design and validation of configurable online aging sensors in nanometer-scale FPGAs. *IEEE Trans. Nanotechnol.* **2013**, *12*, 508–517. [CrossRef]
15. Morales, J.A.; Marc, F.; Bensoussan, A.; Durier, A. Simulation and modelling of long term reliability of digital circuits implemented in FPGA. *Microelectron. Reliab.* **2018**, *88*, 1130–1134. [CrossRef]
16. Jang, B.; Lee, J.K.; Choi, M.; Kim, K.K. On-chip aging prediction circuit in nanometer digital circuits. In Proceedings of the 2014 International SoC Design Conference (ISOCC), Jeju, Korea, 3–6 November 2014; pp. 68–69.
17. Yu, L.; Ren, J.; Lu, X.; Wang, X. NBTI and HCI Aging Prediction and Reliability Screening During Production Test. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2019**, *39*, 3000–3011. [CrossRef]
18. Karimi, N.; Huang, K. Prognosis of NBTI aging using a machine learning scheme. In Proceedings of the 2016 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), Storrs, CT, USA, 19–20 September 2016; pp. 7–10.
19. Vijayan, A.; Chakrabarty, K.; Tahoori, M.B. Machine Learning-Based Aging Analysis. In *Machine Learning in VLSI Computer-Aided Design*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 265–289.
20. Joshi, K.; Mukhopadhyay, S.; Goel, N.; Mahapatra, S. A consistent physical framework for N and P BTI in HKMG MOSFETs. In Proceedings of the 2012 IEEE International Reliability Physics Symposium (IRPS), Anaheim, CA, USA, 15–19 April 2012; pp. 5A.3.1–5A.3.10.
21. Naphade, T.; Goel, N.; Nair, P.; Mahapatra, S. Investigation of stochastic implementation of reaction diffusion (RD) models for NBTI related interface trap generation. In Proceedings of the 2013 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 14–18 April 2013; pp. XT.5.1–XT.5.11.
22. Naouss, M.; Marc, F. FPGA LUT delay degradation due to HCI: Experiment and simulation results. *Microelectron. Reliab.* **2016**, *64*, 31–35. [CrossRef]
23. Khaleghi, B.; Rosing, T. *Reliability Degradation in Nanoscale CMOS: A Review of Modeling, Monitoring, and Mitigation Techniques*; ACM: New York, NY, USA, 2019; pp. 1–23.
24. Gao, Z.; Zhu, J.; Han, R.; Xu, Z.; Ullah, A.; Reviriego, P. Design and implementation of configuration memory SEU-tolerant viterbi decoders in SRAM-based FPGAs. *IEEE Trans. Nanotechnol.* **2019**, *18*, 691–699. [CrossRef]
25. Wong, J.S.; Sedcole, P.; Cheung, P.Y. A transition probability based delay measurement method for arbitrary circuits on FPGAs. In Proceedings of the 2008 International Conference on Field-Programmable Technology, Taipei, Taiwan, 8–10 December 2008; pp. 105–112.
26. Glocker, E.; Chen, Q.; Schlichtmann, U.; Schmitt-Landsiedel, D. Emulation of an ASIC power and temperature monitoring system (eTPMon) for FPGA prototyping. *Microprocess. Microsystems* **2017**, *50*, 90–101. [CrossRef]
27. Prashanth, S.; Sucheta, R.; Vishva, R.; TR, G.K.; Mohan, N. BIST Based Aging Fault Prediction Using Machine Learning. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1715–1722.
28. Zhou, P. *AXI DMA v7.1 LogiCORE IP Product Guide Vivado Design Suite*; Xilinx: San Jose, CA, USA, 14 June 2019.
29. Guo, X.; Burleson, W.; Stan, M. Modeling and experimental demonstration of accelerated self-healing techniques. In Proceedings of the 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 1–5 June 2014; pp. 1–6.
30. Gaskin, T.; Cook, H.; Stirk, W.; Lucas, R.; Goeders, J.; Hutchings, B. Using novel configuration techniques for accelerated FPGA aging. In Proceedings of the 2020 30th International Conference on Field-Programmable Logic and Applications (FPL), Gothenburg, Sweden, 31 August–4 September 2020; pp. 169–175.
31. Li, X.; Qin, J.; Bernstein, J.B. Compact modeling of MOSFET wearout mechanisms for circuit-reliability simulation. *IEEE Trans. Device Mater. Reliab.* **2008**, *8*, 98–121. [CrossRef]
32. Cai, H.; Petit, H.; Naviner, J.F. Reliability aware design of low power continuous-time sigma–delta modulator. *Microelectron. Reliab.* **2011**, *51*, 1449–1453. [CrossRef]
33. Junior, N.G.; Costa, F.J.; Trevisoli, R.; Barraud, S.; Doria, R.T. Influence of interface traps density and temperature variation on the NBTI effect in p-Type junctionless nanowire transistors. *Solid-State Electron.* **2021**, *186*, 108097.

sensors

MDPI

Article

An Unsupervised Condition Monitoring System for Electrode Milling Problems in the Resistance Welding Process

Daniel Ibáñez [1,*], Eduardo Garcia [2], Jesús Soret [1] and Julio Martos [1]

[1] Department of Electronic Engineering, Universidad de Valencia, Campus de Burjassot, 46100 Burjassot, Spain; jesus.soret@uv.es (J.S.); julio.martos@uv.es (J.M.)
[2] Ford Spain, Poligono Industrial Ford S/N, 46440 Almussafes, Spain; egarci75@ford.com
* Correspondence: daniel.ibanez@uv.es; Tel.: +34-961-791-543

Abstract: Resistance spot welding is one of the most widely used metal joining processes in the manufacturing industry, used for structural body manufacturing, railway vehicle construction, electronics manufacturing, battery manufacturing, etc. Due to its wide use, the quality of welded joints is of great importance to the manufacturing industry, as it is critical for ensuring the integrity of finished products, such as car bodies, that withstand high levels of stress. The quality of the welding is influenced both by the programming of the welding and by the good condition of the mechanical part that carries out the welding. These mechanical factors, such as electrode geometry and wear, degrade over time. As the welding points are made, the geometry and properties of the electrodes change, so they undergo a milling process to remove impurities and return them to their initial geometry. Sometimes the milling is deficient, and the electrode continues to wear, causing welding problems such as loose spots and metal spatter. This article presents a method for condition monitoring of the milling process and weld wear based on existing data in real production lines. The use of unsupervised clustering methods is proposed to perform a check by which, using current and resistance data, the electrode wear is grouped. Specifically, a method using multidimensional k-means for the condition monitoring of electrode wear is established. This research gives a real and applicable solution for reducing the quality problems caused by milling defects and electrode wear in the production lines of high-production manufacturing industries, presenting a system for sending alarms based on the behavior of welding variables.

Keywords: resistance spot welding; electrode wear; condition monitoring; milling machine; unsupervised clustering

Citation: Ibáñez, D.; Garcia, E.; Soret, J.; Martos, J. An Unsupervised Condition Monitoring System for Electrode Milling Problems in the Resistance Welding Process. *Sensors* **2022**, *22*, 4311. https://doi.org/10.3390/s22124311

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 13 May 2022
Accepted: 5 June 2022
Published: 7 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Resistance spot welding is one of the most important joining processes in the metallurgy industry due to its efficiency and suitability for automation [1]. Specifically in the automotive industry, modern auto-body assembly needs 7000 to 12,000 spots of welding, and, thus, resistance welding plays a crucial role, representing approximately 90% of the welded joints carried out for body assembly [2].

The welding process can be summarized very simply; the sheet of metal to be welded is placed between water-cooled electrodes and then heat is obtained by passing a large electric current through them for a short period of time [3].

Although this process can be summarized in a very simple way, in reality, there are many factors that affect the achievement of the desired quality. Many programmable parameters affect weld quality. These parameters are given by Joule's law and are the welding time, the current and the resistance, which is related to the pressure achieved by the electrodes [4]. These parameters must be configured to achieve the desired quality and stability over time. In addition, several factors play an important role in the quality of the weld, such as voltage fluctuation, the misalignment of electrodes or loss of electrode

pressure. The shared characteristic of these factors is that they do not change during the lifetime of the welding electrodes and can be better controlled by a better welding controller or machine maintenance [5].

However, another parameter is inherent to the number of welds performed throughout the life of the electrodes: wear. The wear of the electrodes increases as the number of welds increases, modifying both the electrical and thermomechanical properties between the electrodes and the sheets. There are special cases in which this wear is even more pronounced, in particular, in those cases in which the sheets are coated with zinc or when sealer is used between the sheets to be welded. These special cases tend to stick to the copper electrodes, thus, causing a further increase in wear [6].

The heating of the metal can be described according to Joule's law, represented in Equation (1), where Q is the heat generated during welding by passing a current (I) along the metal–metal and metal–electrode resistance (R) over a period of time [7]:

$$Q = I^2Rt \tag{1}$$

In Figure 1, it can be seen schematically, as in resistance spot welding, that three different types of process resistance determine the resistance represented by Joule's law: contact resistors $R3$, $R4$ and $R5$ between sheet metal resistors $R1$ and $R2$. Contact resistance refers to the resistance generated at the interface between the electrode and the metal sheet ($R3$ and $R5$) and the resistance between the metal sheets ($R4$). The resistance of the sheets is determined by the resistivity and the thickness of the metal. Normally, these resistances are higher than the contact resistance between the electrodes, which causes the fusion to begin at the union of the two metals [8].

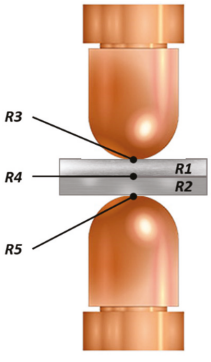


Figure 1. Resistances involved in a resistance spot weld.

From Figure 1, it can also be concluded that good contact between the electrodes and the metal is essential so that the electrode–sheet contact resistance is lower than the metal–metal resistance in such a way that the fusion begins between the metal-to-metal contacts. This bad contact between the electrodes and the metal can be due to different circumstances: misalignment of the electrodes, bad programming of the position of the welding point, dirt on the metals, deformation and wear of the electrodes, etc.

To avoid and correct the wear of the electrodes, a series of milling operations are carried out throughout the electrodes' useful life. Sometimes, these mills fail to reshape the electrode, leaving dirt in the capsule or leaving the electrode with a different shape to the desired one. This causes the contact resistance to vary, and, in addition, an increase in welding quality problems, such as small, deformed or even non-existent nuggets. Due to the increase in resistance and, therefore, the increase in heat, metal ejections can occur during welding, causing a quality and safety problem in the production line [9].

2. Electrode Wear and Milling

Electrode wear is one of the most important issues to research in the resistance welding process. Specifically, studies have focused on determining the factors that influence the appearance of electrode wear. First, Tanaka et al. [10] found that electrode wear could be reduced by increasing the nickel content of the metal foils. In this same line, Rashid et al. [11] demonstrated how a good choice of lubricant coated on the metal sheets could increase the useful life of the electrodes. Similarly, different authors have described the relationship between the different welded materials and the useful life of the electrodes [12–16].

In the same way, the decrease in welding quality caused by the wear of the electrodes has been widely investigated. The reduction in quality is determined by the increase in the diameter of the contact face of the tips [17]. This is because the increase in the diameter of the tip of the electrode, which results in a reduction of the heat generation of the welding joint, causes a decrease in the electrode and is the main reason for the decrease in the size of the nugget [18].

The deposition of the metal coating on the copper electrodes generates a change in the properties of the electrode and, therefore, the wear of the tips [19]. In addition to the reduction of the size of the nugget, the wear of the electrode is of great importance in the presence of weld ejections and other quality defects and can be the cause of 60% of quality problems [20].

Finally, due to the great importance of this defect, different authors have proposed methods for estimating wear or evaluating it. Gauthier et al. [21] and Raoelison et al. [22] demonstrated a method for the numerical modeling of electrode wear which is useful for theoretical estimation but can hardly be applied to real cases where different factors interfere, such as mechanical problems or changes in the production process. Peng et al. [23] proposed the use of images for the analysis of the degradation of the electrodes; the main disadvantage for the application of this system in large production factories is the cost associated with the acquisition of the equipment.

On the other hand, Zhang et al. [24] proposed the use of electrode displacement to determine electrode wear; the discussed method provides a convincing solution but can only be carried out in those welding guns that have an integrated measurement system for electrode displacement, something that is usually lacking in pneumatic welding guns.

Finally, Zhou et al. [25] presented a method based on the analysis of dynamic resistance during welding to determine wear. The main disadvantage of this study is that it assumed that the dynamic resistance trend depends only on the wear of the electrodes when, actually, this variation can depend on different factors, such as the final quality of the welding point, as stated by Zhao et al. [26], or the temperature and pressure, as stated by Whang et al. [27], among many other factors.

All these proposed methods were based on the premise that, after performing the milling, the electrodes return to their original geometric state. On multiple occasions, due to mechanical problems of both the welding gun and the milling machine, such as blade wear or issues with milling, as shown in Figure 2, the restoration of the geometry does not occur.

(a) (b)

Figure 2. Typical defects of milling machines. (**a**) Milling machine covered by the chips from the electrodes. (**b**) Dull blade.

Therefore, the objective of this research is not to propose a system that only determines the wear of the electrode, but one which determines the milling problems that may occur during the production process; that is, the main objective of this research is to avoid the quality problems caused by the wear of the electrodes.

3. Materials and Methods

For the creation of the milling problem detection system, it was essential to be able to relate a real variable to an existing defect; this variable had to be acquired and treated to then proceed to the analysis and the elaboration of a data analysis system for evaluation.

Specifically, due to its existing relationship, the use of the measurement of the electrode resistance is proposed for subsequent preprocessing with normalization and feature extraction to later carry out an unsupervised classification method. This allows the setting of detection thresholds based on the behavior of the resistance data.

3.1. Electrode Resistance Measurement Method

Electrode wear is one of the essential external factors that determines the stability of weld joints in the resistance welding process.

To avoid these quality problems, after a certain number of welding points, a shaping of the tips is carried out by means of a milling machine. This process can be automatic or manual depending on the type of production line. Sometimes, due to a malfunction of the milling machine, such as an issue with the cutter, a force problem in the welding gun, poor positioning of the milling machine, etc., the electrodes are not well shaped. This is a problem since, until the next milling or replacement of electrodes, they will continue to function with inadequate wear, which can cause serious quality problems.

Figure 3 shows the real differences between electrodes after adequate milling and those after defective milling. The electrodes in Figure 3a correspond to 16 mm type F electrodes, according to DIN EN ISO 5821, before being milled for the first time. Figure 3b shows some electrodes that, after executing a certain number of welding points, were milled and returned to their original geometry. Finally, comparing Figure 3b,c, a clear example of poor milling can be seen. In Figure 3c, the active face of the electrodes has been partially cleaned, showing the dirt that generates quality problems.

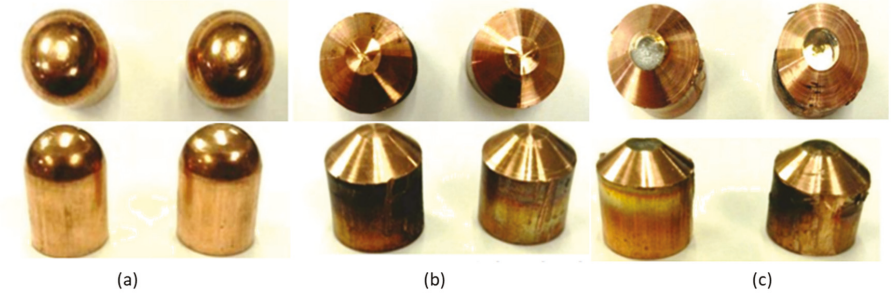


Figure 3. Evolution of the state of the electrodes. (**a**) Type F electrode before milling. (**b**) Type F electrode after good milling. (**c**) F-type electrode after poor milling.

Due to this uncertainty regarding the good milling of the electrodes, a method was established to measure the resistance after each milling is performed, acquiring the voltage and the average current measured between the short-cut electrodes, as shown in Figure 4.

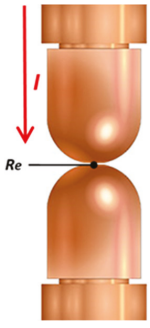


Figure 4. Schema—electrode resistance contact (Re) measurement.

This check is carried out at a constant primary voltage so that when there is a change in the contact resistance of the electrodes (Re) due to the wear of the electrodes, the voltage measured at the electrodes and the current vary according to Ohm's law.

3.2. Data Adquisition

For this article, 650 welding guns located in a real production line were analyzed, as well as a total of 100 millings carried out for each of the electrodes. The welding controls used for the study were BOS6000 with medium-frequency transformers. Regarding the welding guns, the analysis was carried out with no differences between pneumatic guns and servo guns. Similarly, two different welding electrodes were used for the study, 6 mm and 8 mm tip face electrodes, but, at the time of analysis, this difference was considered insignificant for the detection of electrode wear.

In relation to the type of milling machine and electrode, milling machines with an average speed greater than 290 min^{-1} and 25 Nm of torque were used to reset the geometry of the electrode, capable of restoring the geometry of the electrodes according to DIN EN ISO 5821 F1-16-20 [28].

For data acquisition, a pipeline system was implemented between the welding controller and the database through the ELK stack [29]. In this way, a protocol was established to send the welding data to the database every time a welding point occurred, which allowed real-time data analysis, both for machine failure detection and, in this case, for the performance of predictive analysis of weld quality.

For our case, the data acquisition protocol was established, as shown in Figure 5. In the first place, once the maximum number of the welding joints that an electrode could make was reached, keeping the welding quality constant, the electrodes were sent for milling. When the milling was finished, the electrodes were short-circuited by passing a current at constant voltage (PHA method). Finally, the data were stored in the database for further analysis.

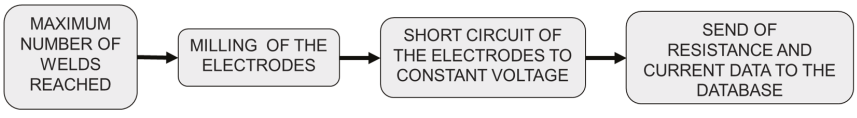


Figure 5. Data acquisition protocol.

3.3. Proprocessing and Feature Extraction

Once the necessary programming was carried out in the welding lines, all the data of the 650 welding guns were stored in the database.

In Figure 6, the data for two different welding guns are shown; it can be seen that the average value of the resistance for each of the cases was different. The difference observed was due to the characteristics of each of the guns, which depended on where the terminals

of the voltage probe were located; they affected not only the resistance of the electrodes but also the resistance of the welding arm. The data were always analyzed as normalized data to eliminate this difference from the analysis. Therefore, the z-score normalization, shown in Equation (2), was used. This normalization based on the mean and the standard deviation allowed the reduction of variations if high current and resistance data series were added to the analysis [30].

$$x' = \frac{x - \overline{x}}{\sigma} \tag{2}$$

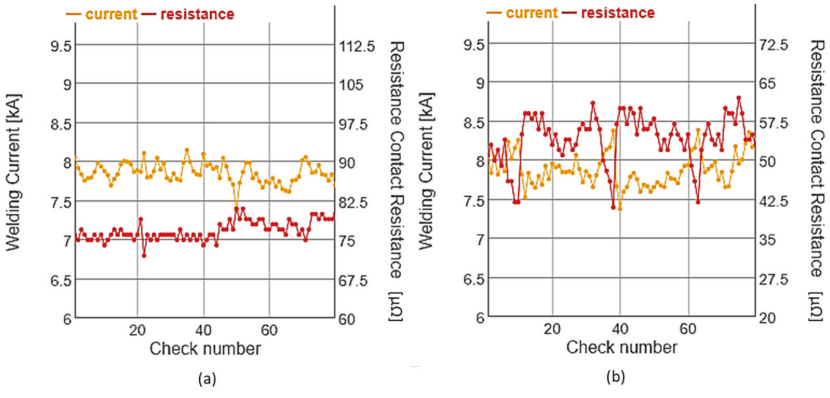


Figure 6. Example of current and resistance data acquired for a weld control. (**a**) Example of the data acquired for a welding gun with correct milling; (**b**) Example of the data acquired for a welding gun with poor milling.

Similarly, in this preprocessing stage, the data were subjected to data cleaning. First, the existing zeros in the time series were removed since those values were meaningless. This is because, when problems appeared when carrying out the check or in the voltage and current measurement probes, a zero was stored in the database. After eliminating the zeros, the atypical data of the time series were calculated, and the outliers were eliminated, for which the expressed formula in Equation (3) was used.

$$\begin{aligned} Max &= Q_3 + 1.5IQR \\ Min &= Q_1 - 1.5IQR \end{aligned} \tag{3}$$

Once this signal was filtered, the feature extraction process was carried out. Feature extraction in machine learning is a process of extracting significant attributes of the data. Feature extraction allows the height of dimensions of a series of data to be reduced to a smaller number of dimensions through unique mapping techniques [31].

For this study, the time series of both resistance and current were reduced to five statistical variables, which allowed the reduction of the dimensions by eighty times for each signal. The calculated variables were:

- The coefficient of variation (CV): the ratio of the standard deviation to the mean;
- Quartile Q_1;
- Quartile Q_3;
- Inter-decile range (IDR): the difference between D9 and D1;
- Median.

As there were two summary signals, in total we had 10 statistical variables as an input dataset for each welding control. The input dataset for the k-mean algorithm was a 650 × 10 array of values. Finally, before proceeding to k-means, the input dataset was standardized to obtain a more precise result in the next section.

3.4. K-Means Clustering

In this research, the use of unsupervised clustering using k-means was proposed, Algorithm 1. In general, for this method, the optimal number of clusters for the existing amount of data to be processed is selected first. This parameter represents the number of desired groupings.

Algorithm 1: K-means Clustering.

Input:
$X = \{x_1, x_2, \ldots \ldots, x_n\}$ (input data)
k (number of clusters)
Output:
$C = \{c_1, c_2, \ldots \ldots, c_k\}$ (set of cluster centroids)
Initialization:
for each $c_i \in C$ do:
$c_i \leftarrow x_j \in X$ (random selection)
while: Convergence or max iteration reached
for each $x_i \in X$ **do**:
minDist $\leftarrow$ *minDistance*(x_i, c_j) $j \in (1 \ldots k)$;
(based on Euclidian distance $\frac{1}{n}\sum(\min j\, d^2\left(x_i, c_j\right))$ for $i = 1$ to n)
UpdateCluster(c_i)

Based on the dataset, the k-means groups them in the programmed number of clusters k, assigning them to the closest centroid. Finally, the algorithm returns both the cluster and the respective centroid. Starting from an initial, non-optimized grouping, the algorithm relocates each point to the nearest new center. It then updates the centers of each cluster based on the mean of the points, repeating this relocation and updating the process until the convergence criteria are satisfied; this process is summarized in the flowchart of Figure 7 [32–34].

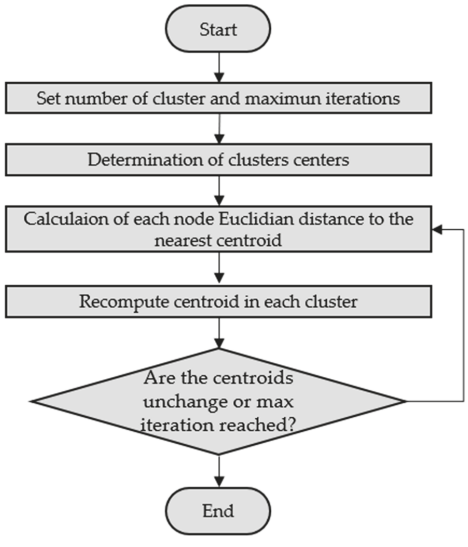


Figure 7. K-means flowchart.

One of the main advantages of using k-means and unsupervised learning is that it is not necessary to have labeled data. In this study, the population of equipment analyzed was large and varied, which is why it was difficult and inaccurate to label each series of data with the current state of the machine. In this way, it was not necessary to know the characteristics of each of the possible faults that may occur in the milling or in the

electrodes, but rather the k-means algorithm, based on the behaviors, assigned each series of data to each cluster.

The purpose of this analysis was to detect any variation in the milling process through its influence on the k-means clustering algorithm. In this case, three different data behavior clusters were expected, and we aimed to establish three machine status criteria: alarm, pre-alarm and good status.

As previously mentioned, the ten statistical variables calculated for the simplification of the model were used as an input dataset for the k-means clustering algorithm. There are different techniques to determine the optimal number of clusters, such as silhouette width, AIC [35] and BIC [36] within the sum of the square (WSS) [37] and NbClust [38]. In this investigation, given that the performance of the AIC and BIC techniques decreases as the dimensionality of the data increases [39] and that the NbClust technique has higher precision than the WSS technique, the optimal number of clusters was identified by the NbClust technique. Specifically, as can be seen in Figure 7, the NbClust function for the input dataset discussed above gave the optimal cluster number for the k-means of the three clusters.

In Figure 8a, the result of the average silhouette technique for choosing the optimal number of clusters is shown; it can be seen that the results for two and three clusters were very close, although the test showed that two was the optimal number of clusters. The same is observable in Figure 8c; although the values of two and three were similar, this technique stated that the optimal choice was two clusters. On the other hand, using the elbow method, as shown in Figure 8b, it was observed that the optimal number of clusters was between three and four clusters. Finally, in Figure 8d, corresponding to the results of the NbClust method, it can be seen that the number of optimal clusters was between two and four, obtaining the highest result for three clusters. Therefore, based on these four analyses and taking into account the greater reliability of the NbClust method, it was established that the optimal number of clusters in this study for the input dataset was three.

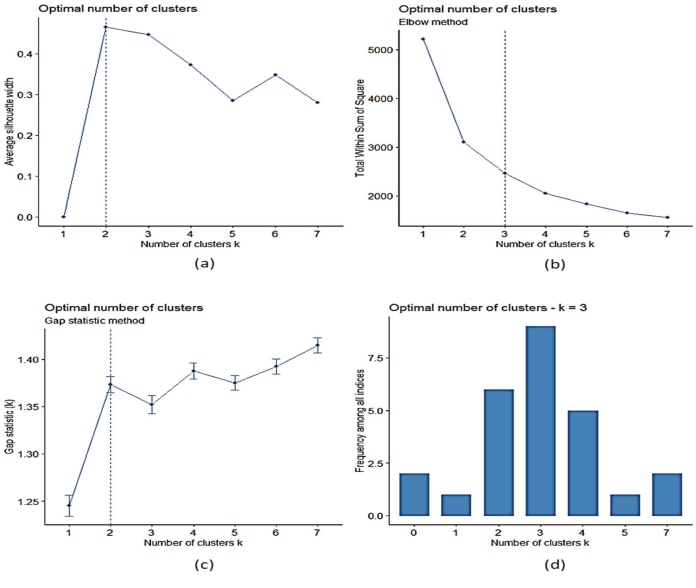


Figure 8. Determination of the optimal number of clusters, (**a**) average silhouette, (**b**) WSS and elbow technique, (**c**) gap statistic method, (**d**) NbClust.

4. Evaluation and Results

Throughout the previous section, the methodology used, the signals that were analyzed and their dimensional conversion into statistical variables were shown, ending with the method used for clustering and the optimal number of clusters for the proposed dataset.

This section shows the results obtained after using k-means for the grouping of the input dataset. First, the statistical data of each cluster generated were analyzed to determine the behavior corresponding to each cluster.

Table 1 shows the average distance between the points per cluster pair and the distance between the centers of each of the clusters. Several conclusions can be drawn from these values; the distance between centroids was greater between cluster 2 and cluster 3, so cluster 1 could be considered as the central cluster of data deviation, establishing itself as the pre-alarm cluster. Similarly, it was observed that the distance between cluster 1 and cluster 2 was greater than the distance between cluster 2 and cluster 3. This suggests that, due to the dispersion of the data, cluster 2 could be the cluster of points in alarm state.

Table 1. Matrix of separation values between all pairs of clusters and distance between centroids.

Cluster	1		2		3	
	Separation	**Centroids**	**Separation**	**Centroids**	**Separation**	**Centroids**
1	0.00	0.00	0.98	5.66	0.48	3.11
2	0.98	5.66	0.00	0.00	3.70	8.10
3	0.481	3.11	3.70	8.10	0.00	0.00

Table 2 shows the centroids obtained by k-means for each of the input variables; these centroids are the ones that were used to assign the membership of the checks to each cluster and, therefore, their alarm status.

Table 2. Cluster centroids for each dimension.

Cluster	C CV	C Q_1	R CV	R Q_1	C Q_3	R Q_3	IDR R	C Q_2	R Q_2	IDR C	DIM 1	DIM 2
1	0.71	0.64	0.48	0.65	−0.66	−0.62	0.83	−0.16	0.25	0.82	−1.58	−0.18
2	2.23	2.19	2.11	2.39	−2.39	−2.15	1.96	2.42	−2.61	1.83	1.86	0.67
3	−0.60	−0.55	−0.45	−0.57	0.58	0.54	−0.65	−0.11	0.07	−0.63	6.69	−2.26

To simplify the cluster plotting process for analysis, these centroids were dimensionally reduced from being ten dimensions, one for each input variable, to two dimensions. These two dimensions were obtained by means of PCA [40]. In Table 2, this reduction in dimensions can be observed with the centroids for dimension 1 (DIM 1) and dimension 2 (DIM 2).

To continue with the analysis of the results, the graphs in Figure 9 were analyzed. In this figure, the clusters are represented after their dimensional reduction to two unique dimensions, DIM 1 and DIM2, in order to plot a simpler graph. In the figure, four graphs are represented; two of them show the density distribution for each dimensional reduction. With the help of these two graphs, it can be concluded that, in cluster 3, there were data of those guns with a more stable milling and, therefore, they were correct. This can be confirmed since, observing the distributions of cluster 3 in both dimensions, it can be seen that there was a lower dispersion and a greater normality compared to the other clusters.

In the same way, following the same reasoning as for cluster 3, it was established that cluster 1 represents the millings that begin to be deficient, while cluster 2 groups the deficient millings that start to create quality problems in the welding points due to excessive wear of the electrodes.

Finally, Figure 10 shows the current graphs corresponding to each of the clusters. In Figure 10, three current curves grouped in cluster 3 are shown, which correspond to correct

operation; if compared with Figure 10a it can be observed that the curves of both graphs have a low dispersion and a stable behavior.

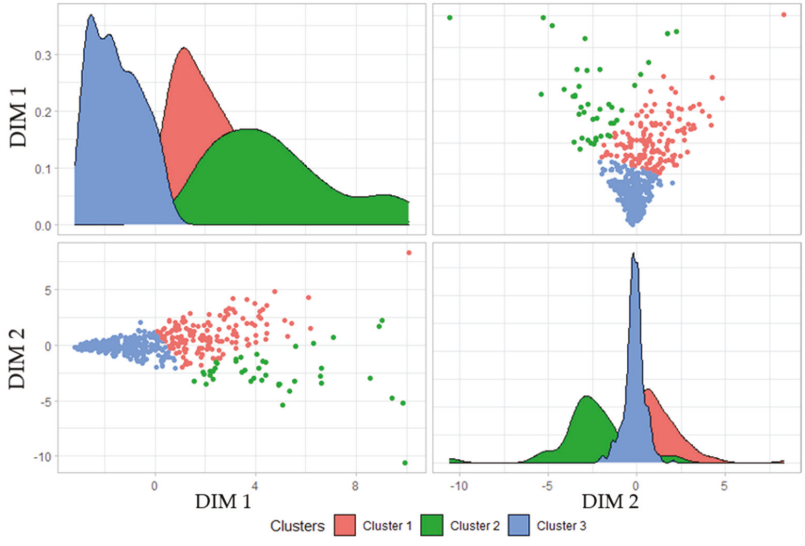


Figure 9. Result with dimensional reduction of the clustering for the input dataset.

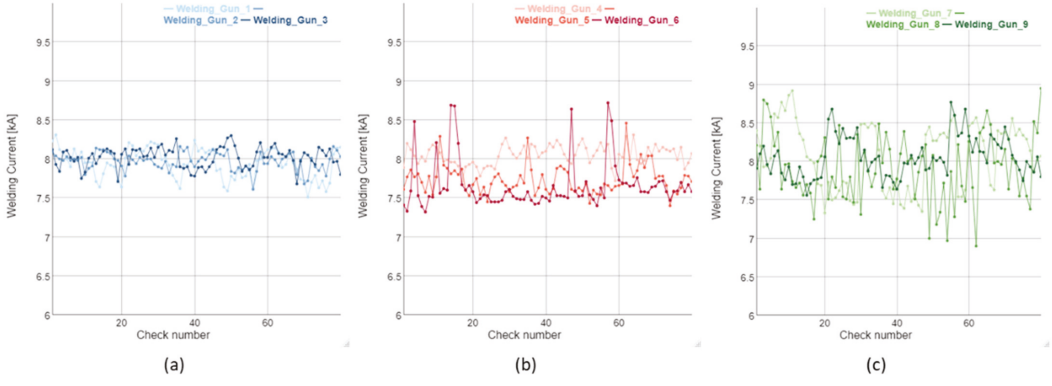


Figure 10. Examples of current measured depending on the cluster for different welding guns. (**a**) Cluster 3, determined as correct operation, (**b**) Cluster 1, determined as pre-alarm, (**c**) cluster 2, determined as alarm.

Figure 10b,c shows the current curves for clusters 1 and 2, respectively. From their analysis, it can be concluded that, as the data are assigned from pre-alarm cluster 1 to alarm cluster 2, the curves begin to show greater variance instability, which is an unequivocal sign that the electrodes presented a problem in milling and, therefore, increased wear, which will inevitably turn into quality problems at the weld point.

5. Application of the System for Real-Time Detection

This research was not only focused on finding a method that allows the detection of milling problems. The high production rate of manufacturing factories makes it essential that production defects are evaluated in real time. This allows the reduction of costs caused by having to repair products manufactured with poor quality since early detection can reduce the number of poor-quality products manufactured.

The clustering method for the detection of milling problems and electrode wear groups the behavior of the data series in three differentiated clusters: correct operation, pre-alarm and alarm. These three clusters, therefore, allow an algorithm to evaluate and label the status of each of the welding guns in a factory. The real-time detection system is designed to analyze each welding gun in particular and send the operators in charge of that welding line the alarm so that the defect and its possible consequences can be repaired.

As mentioned in the previous sections, a communication structure is necessary between the welding equipment and the database so that the data from all the welding equipment is accessible from the data analysis software. The system for detecting milling problems and electrode wear first collects resistance and current data from each of the welding equipment, labeling those controls that do not have enough data due to communication problems. Next, the extracted data are normalized, as described in the previous sections, and the dimensional reduction of input variables is carried out.

Once the reduction of the time series to the ten input variables has been carried out, the cluster each one of the analyzed pieces of welding equipment belongs to is determined. The assignment of each cluster is carried out by calculating the distance between each point with respect to the centroids of each of the clusters.

The assignment of each of the clusters determined after measuring the distance to each of them allows each piece of the welding equipment to be labeled according to its status in such a way that the welding equipment that is assigned to cluster 3 presents correct operation, and those in cluster 2 are in alarm and, therefore, require corrective action to be carried out.

Finally, once it has been determined that the welding guns have a behavior typical of electrodes with high wear, the alarm dispatch system is carried out to the production lines. In this case, a publish/subscribe protocol based on AMQP is established [41]. This protocol allows the sending of messages in specific queues. In this case, queues managed by RabbitMQ are used, which allows the sending of alarms through Webex to those in charge of repairing the conflicting equipment. The system is like the one proposed by García and Montes [42] for the acquisition of data from PLC in real time in factories, but, in this case, it is not based on data stored in a PLC but rather the welding equipment itself stores the data through Logstash, making it accessible from the data collector. In short, as described in Figure 11, the proposed system collects data directly from the real production lines and, after data processing, can label defects and send alarms for the repair and correction of quality problems produced.

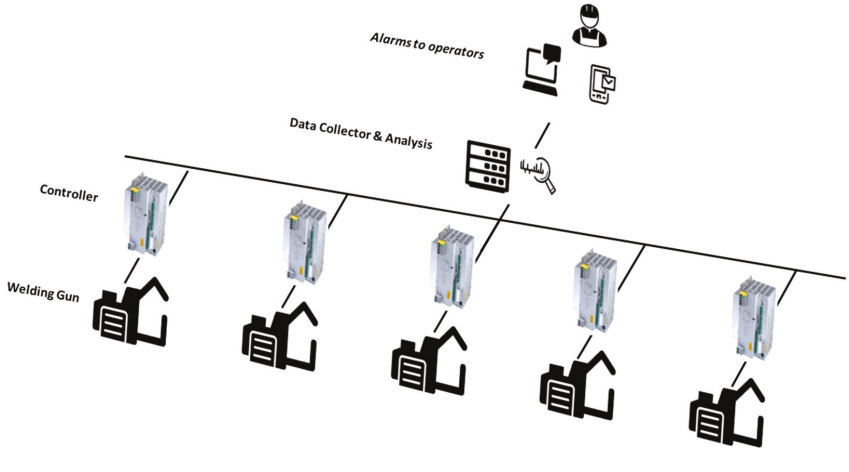


Figure 11. Real-time data collection schema.

6. Conclusions

This research presents a novel method for the detection of milling problems and electrode wear using unsupervised clustering methods. Throughout this article, the relationship between the serial time data of resistance and the variation of the mechanical properties of the electrodes was described.

Despite working with time series, feature extraction was carried out to reduce dimensionality, which allowed the reduction of the number of input inputs of the clustering algorithm. This also allowed the input data to be scaled so that they were not influenced by the resistance differences existing in each welding gun.

The main advances obtained from this research were the following:

- A method for detecting the relationship between welding variables and milling state;
- An alarm system, where pre-alarm status and correct operation are established according to the output of the clustering algorithm;
- A system for the collection of data in a welding line that allows the realization of data analysis in real time, both for this investigation and for future investigations.

Despite the advances described, the system is still not capable of differentiating between types of fault. Different mechanical factors influence milling problems, such as worn blades, transformer secondary circuit problems, etc. The objective of future work in this investigation should go from the cataloging of faults as alarm, pre-alarm and correct status to a fault labeling system based on behavior pattern. In the same way, throughout this investigation, unsupervised learning methods were used due to the characteristics of the sample, but, in future works, we expect to continue in the line of experimentation with other analysis methods that could improve the detection system.

Author Contributions: Conceptualization, D.I., E.G. and J.S.; Formal analysis, D.I.; Investigation, D.I., E.G., J.S. and J.M.; Methodology, J.S. and J.M.; Project administration, E.G. and J.M.; Supervision, E.G. and J.S.; Validation, D.I. and J.M.; Visualization, D.I. and J.M.; Writing—original draft, D.I.; Writing—review & editing, E.G., J.S. and J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Ford España S.A., and, in particular, the Almussafes Factory, for their support in the investigation. Likewise, the authors express their greatest gratitude to the "Fundación para el desarrollo y la innovación" (FDI), together with the Generalitat Valenciana, for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hou, Z.; Kim, I.-S.; Wang, Y.; Li, C.; Chen, C. Finite element analysis for the mechanical features of resistance spot welding process. *J. Mater. Process. Technol.* **2007**, *185*, 160–165. [CrossRef]
2. Huh, H.; Kang, W. Electrothermal analysis of electric resistance spot welding processes by a 3-D finite element method. *J. Mater. Process. Technol.* **1997**, *63*, 672–677. [CrossRef]
3. Özyürek, D. An effect of weld current and weld atmosphere on the resistance spot weld ability of 304L austenitic stainless steel. *Mater. Des.* **2008**, *29*, 597–603. [CrossRef]
4. Aravinthan, A.; Nachimani, C. Analysis of Spot Weld Growth on Mild and Stainless Steel. *Weld. J.* **2011**, *90*, 143–147.
5. Tang, H.; Hou, W.; Hu, S.J.; Zhang, H.Y.; Feng, Z.; Kimchi, M. Influence of welding machine mechanical characteristics on the resistance spot welding process and weld quality. *Weld. J.* **2003**, *82*, 116-S–124-S.
6. Zhang, J.; Zhang, P.X.; Xu, X.J. A Model for Predicting the Wear Degree of Electrode Tip. *Appl. Mech. Mater.* **2014**, *574*, 292–297. [CrossRef]
7. Jeffus, L. *Welding Principals and Applications*, 4th ed.; Thomson Learning: Philadelphia, PA, USA, 1999; pp. 678–681.

8. Del Vecchio, E.J. (Ed.) *Resistance Welding Manual*; Resistance Welder Manufacturers' Association: Philadelphia, PA, USA, 1956; Volume 1.
9. Zhang, X.; Chen, G.; Zhang, Y. Characteristics of electrode wear in resistance spot welding dual-phase steels. *Mater. Des.* **2008**, *29*, 279–283. [CrossRef]
10. Tanaka, Y.; Sakaguchi, M.; Shirasawa, H. Electrode life in resistance spot welding of zinc plated steel sheets. *Int. J. Mater. Prod. Technol.* **1987**, *2*, 64–74.
11. Rashid, M.; Fukumoto, S.; Medley, J.B.; Villafuerte, J.; Zhou, Y. Influence of lubricants on electrode life in resistance spot welding of aluminum alloys. *Weld. J.* **2007**, *86*, 62-s.
12. Kondo, M.; Konishi, T.; Nomura, K.; Kokawa, H. Degradation mechanism of electrode tip during alternate resistance spot welding of zinc-coated galvannealed and uncoated steel sheets. *Weld. Int.* **2013**, *27*, 770–778. [CrossRef]
13. Gould, J.E.; Kimchi, M.; Campbell, D.H. *Weldability and Electrode Wear Characteristics of Hot-Dip Galvanized Steel with and without a Ferrophos Containing Primer*; Report No. 880370; SAE: Detroit, MI, USA, 1988.
14. Saito, T.; Takahashi, Y.; Nishi, T. Electrode Tip Life in Resistance Spot Welding of Zinc and Zinc Alloy Coated Sheet Steels. *Nippon. Steel Tech. Rep.* **1988**, *37*, 24–30.
15. Athi, N.; Cullen, J.; Al-Jader, M.; Wylie, S.; Al-Shamma'A, A.; Shaw, A.; Hyde, M. Experimental and Theoretical Investigations to the Effects of Zinc Coatings and Splash on Electrode Cap Wear. *Measurement* **2009**, *42*, 944–953. [CrossRef]
16. De, A.; Dorn, L.; Gupta, O. Analysis and Optimisation of Electrode Life for Conventional and Compound Tip Electrodes During Resistance Spot Welding of Electrogalvanised Steels. *Sci. Technol. Weld. Join.* **2000**, *5*, 49–57. [CrossRef]
17. Fukumoto, S.; Lum, I.; Biro, E.; Boomer, D.R.; Zhou, Y. Effects of electrode degradation on electrode life in resistance spot welding of aluminum alloy 5182. *Weld. J.* **2003**, *82*, 307–312.
18. Wang, B.; Hua, L.; Wang, X.; Song, Y.; Liu, Y. Effects of electrode tip morphology on resistance spot welding quality of dp590 dual-phase steel. *Int. J. Adv. Manuf. Technol.* **2015**, *83*, 1917–1926. [CrossRef]
19. Chan, K.R. Weldability and Degradation Study of Coated Electrodes for Resistance Spot Welding. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2005.
20. Xia, Y.-J.; Su, Z.-W.; Li, Y.-B.; Zhou, L.; Shen, Y. Online quantitative evaluation of expulsion in resistance spot welding. *J. Manuf. Process.* **2019**, *46*, 34–43. [CrossRef]
21. Gauthier, E.; Carron, D.; Rogeon, P.; Pilvin, P.; Pouvreau, C.; Lety, T.; Primaux, F. Numerical Modeling of Electrode Degradation During Resistance Spot Welding Using CuCrZr Electrodes. *J. Mater. Eng. Perform.* **2014**, *23*, 1593–1599. [CrossRef]
22. Raoelison, R.; Fuentes, A.; Pouvreau, C.; Rogeon, P.; Carré, P.; Dechalotte, F. Modeling and numerical simulation of the resistance spot welding of zinc coated steel sheets using rounded tip electrode: Analysis of required conditions. *Appl. Math. Model.* **2014**, *38*, 2505–2521. [CrossRef]
23. Peng, J.; Fukumoto, S.; Brown, L.; Zhou, N. Image analysis of electrode degradation in resistance spot welding of aluminium. *Sci. Technol. Weld. Join.* **2004**, *9*, 331–336. [CrossRef]
24. Zhang, Y.S.; Wang, H.; Chen, G.L.; Zhang, X.Q. Monitoring and intelligent control of electrode wear based on a measured electrode displacement curve in resistance spot welding. *Meas. Sci. Technol.* **2007**, *18*, 867–876. [CrossRef]
25. Zhou, L.; Li, T.; Zheng, W.; Zhang, Z.; Lei, Z.; Wu, L.; Zhu, S.; Wang, W. Online monitoring of resistance spot welding electrode wear state based on dynamic resistance. *J. Intell. Manuf.* **2020**, *33*, 91–101. [CrossRef]
26. Zhao, D.; Bezgans, Y.; Wang, Y.; Du, W.; Vdonin, N. Research on the correlation between dynamic resistance and quality estimation of resistance spot welding. *Measurement* **2021**, *168*, 108299. [CrossRef]
27. Wang, S.C.; Wei, P.-S. Modeling Dynamic Electrical Resistance During Resistance Spot Welding. *J. Heat Transf.* **2000**, *123*, 576–585. [CrossRef]
28. *DIN EN ISO 5821:2010-04*; Resistance Welding—Spot Welding Electrode Caps (ISO 5821:2009). German Version EN ISO 5821:2009; ISO: Geneva, Switzerland, 2009.
29. Michael, M.; Davvid, S. The Rise of Elastic Stack. 2016. Available online: https://www.researchgate.net/publication/309732494_The_Rise_of_Elastic_Stack?channel=doi&linkId=5820655c08ae40da2cb4e19a&showFulltext=true (accessed on 20 April 2022).
30. Mohamad, I.B.; Usman, D. Standardization and Its Effects on K-Means Clustering Algorithm. *Res. J. Appl. Sci. Eng. Technol.* **2013**, *6*, 3299–3303. [CrossRef]
31. Lakshmanan, M.; Karnan, H.; Natarajan, S. *Smart Diagnosis of Cardiac Arrhythmias Using Optimal Feature Rank Score Algorithm for Solar Based Energy Storage ECG Acquisition System*; Academic Press: Cambridge, MA, USA, 2020. [CrossRef]
32. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965 and 27 December 1965–7 January 1966; Statistical Laboratory of the University of California: Berkeley, CA, USA, 1967.
33. Steinley, D. K-means clustering: A half-century synthesis. *Br. J. Math. Stat. Psychol.* **2006**, *59 Pt 1*, 1–34. [CrossRef] [PubMed]
34. Swana, E.; Doorsamy, W. An Unsupervised Learning Approach to Condition Assessment on a Wound-Rotor Induction Generator. *Energies* **2021**, *14*, 602. [CrossRef]
35. Akaike, H. Factor analysis and AIC. *Psychometrika* **1987**, *52*, 371–386. [CrossRef]
36. Burnham, K.P.; Anderson, D.R. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]

37. Amruthnath, N.; Gupta, T. Fault Class Prediction in Unsupervised Learning using Model-Based Clustering Approach. In Proceedings of the 2018 International Conference on Information and Computer Technologies (ICICT), Libertad City, Ecuador, 10–12 January 2018; pp. 5–12.
38. Malika, C.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **2014**, *61*, 1–36.
39. Broman, K.W.; Speed, T.P. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2002**, *64*, 641–656. [CrossRef]
40. Salem, N.; Hussein, S. Data dimensional reduction and principal components analysis. *Procedia Comput. Sci.* **2019**, *163*, 292–299. [CrossRef]
41. AMQP: Advanced Message Queuing Protocol, Version 0.8. AMQP Working Group Protocol Specification. 2006. Available online: https://www.rabbitmq.com/resources/specs/amqp0-8.pdf (accessed on 20 April 2022).
42. Garcia, E.; Montes, N. Mini-term, a novel paradigm for fault detection. *IFAC-PapersOnLine* **2019**, *52*, 165–170. [CrossRef]

sensors

MDPI

Article

Rolling Bearing Fault Diagnosis Based on Markov Transition Field and Residual Network

Jialin Yan [1,2], Jiangming Kan [1,2] and Haifeng Luo [1,2,*]

[1] School of Technology, Beijing Forestry University, Beijing 100083, China; yanjialin@bjfu.edu.cn (J.Y.); kanjm@bjfu.edu.cn (J.K.)
[2] Key Laboratory of State Forestry Administration on Forestry Equipment and Automation, Beijing 100083, China
* Correspondence: luohaifeng@bjfu.edu.cn

Abstract: Data-driven rolling-bearing fault diagnosis methods are mostly based on deep-learning models, and their multilayer nonlinear mapping capability can improve the accuracy of intelligent fault diagnosis. However, problems such as gradient disappearance occur as the number of network layers increases. Moreover, directly taking the raw vibration signals of rolling bearings as the network input results in incomplete feature extraction. In order to efficiently represent the state characteristics of vibration signals in image form and improve the feature learning capability of the network, this paper proposes fault diagnosis model MTF-ResNet based on a Markov transition field and deep residual network. First, the data of raw vibration signals are augmented by using a sliding window. Then, vibration signal samples are converted into two-dimensional images by MTF, which retains the time dependence and frequency structure of time-series signals, and a deep residual neural network is established to perform feature extraction, and identify the severity and location of the bearing faults through image classification. Lastly, experiments were conducted on a bearing dataset to verify the effectiveness and superiority of the MTF-ResNet model. Features learned by the model are visualized by t-SNE, and experimental results indicate that MTF-ResNet showed better average accuracy compared with several widely used diagnostic methods.

Keywords: intelligent fault diagnosis; Markov transition field; residual network

Citation: Yan, J.; Kan, J.; Luo, H. Rolling Bearing Fault Diagnosis Based on Markov Transition Field and Residual Network. *Sensors* **2022**, *22*, 3936. https://doi.org/10.3390/s22103936

Academic Editors: Dong Wang, Shilong Sun and Changqing Shen

Received: 18 April 2022
Accepted: 21 May 2022
Published: 23 May 2022

1. Introduction

Rolling bearings are critical components in rotating machinery, and their operating conditions under various loads directly impact their performance, stability, and endurance. More specifically, rolling bearings are vital in mechanical equipment. To maintain the normal operation of mechanical equipment, it is necessary to monitor the vibration signals generated by the rotating mechanism in real time [1]. Many scholars extensively studied the fault detection and diagnosis of rolling bearings [2–4]. The traditional manual diagnostic can no longer adapt to the large-capacity, diverse, and high-speed data in the current mechanical field, which leads to poor diagnosis capability and generalization performance in the face of massive amounts of mechanical equipment data with alternating multiple working conditions and the serious coupling of fault information [5].

The diagnosis of rolling bearings generally consists of two stages: feature extraction and classification. Signal processing approaches that are widely employed to extract features from a raw signal include short-time Fourier transform (STFT) [6], wavelet transform (WT) [7], and empirical mode decomposition (EMD) [8]. However, traditional fault diagnosis methods rely heavily on manual feature engineering and expert knowledge, and the process is time-consuming and laborious. In addition, when extracted features are insufficient, the accuracy of fault diagnosis is greatly reduced, which is not conducive to the diagnostic tasks of massive amounts of industrial data. In the past decade, machine-learning theories and statistical inference techniques have been widely applied to identify

bearing faults, such as Bayesian networks [9], artificial neural networks (ANNs) [10], support vector machines (SVMs) [11], and k-nearest neighbor [12]. Despite the effectiveness of the above-mentioned methods, shallow networks are restricted in their capacity to represent complicated functions with limited samples; thus, they lack the ability to diagnose the faults of complex and high-dimensional signals.

In recent years, deep-learning models have grown in popularity in the field of machine learning, which uses the deep network structure to achieve more efficient and reliable feature extraction. Deep learning disposes of the dependence on manually extracting features and expert experience, which has achieved breakthroughs in many pattern recognition tasks such as natural-language processing [13], automatic speech recognition [14], and computer vision [15]. The application of deep-learning models in fault diagnosis and health monitoring is flourishing [16,17]. Shao et al. [18] proposed a new deep belief network, which was optimized with the particle swarm algorithm, and verified the robustness of the model. Wen et al. [19] developed a novel DTL model for fault diagnosis that extracted features with a three-layer sparse autoencoder and achieved high prediction accuracy. Jiang et al. [20] constructed a deep recurrent neural network with an adaptive learning rate for the fault diagnosis of bearings, and results confirmed the effectiveness of the method. Hasan et al. [21] proposed an explainable AI-based fault diagnosis model and incorporated explainability to the feature selection process. Within the deep-learning framework, convolutional neural networks, as an end-to-end learning model with powerful feature extraction capability, have received more attention in fault diagnosis. Chen et al. [22] developed bearing discrimination patterns on the basis of the cyclic spectral coherence (CSCoh) maps of vibration signals and established a CNN model to learn high-level features. Guo et al. [23] proposed a new method named DCTLN for transfer fault diagnosis tasks, and verified the effectiveness of the model by experiments. Jia et al. [24] proposed a DNCNN to address imbalanced classification problems in fault diagnosis. In some scenarios, raw one-dimensional signals are converted into two-dimensional gray images with pixels fulfilled by data stacking [25,26]. However, these methods may contain limited feature information because spatial correlation in a raw vibration sequence can be corrupted. Although there are a few commonly used image representation approaches based on time–frequency principles, such as short-time Fourier transform (STFT) [6] and wavelet packet transform (WPT) [27], short-time Fourier transform is not suitable for handling nonstationary signals such as mechanical fault signals, and the determination of the number of decomposition layers for wavelet packets usually relies heavily on expert knowledge. Therefore, a new image encoding method called Markov transition field (MTF) was introduced [28] that preserves complete time-domain information by representing Markov transition probabilities, and converts that information into two-dimensional images. In addition, despite the great success of deep convolutional neural networks, degradation problems such as gradient disappearance or explosion can occur as the number of layers increases. To address the issue mentioned above, He et al. [29] proposed residual networks that have achieved excellent performance on various machine-learning tasks.

In order to efficiently represent the state characteristics of vibration signals in image form and improve the feature learning capability of the network, a new intelligent bearing fault diagnosis method (MTF-ResNet) is proposed in this paper. The main contributions of this paper are summarized as follows.

1. A novel two-step fault diagnosis method is proposed that converts raw vibration signals into images through the Markov transition field, and adopts the residual network for feature extraction and fault identification.
2. The signal-to-image conversion preserves the time dependence of the raw vibration signals and retains sufficient temporal features without setting parameters involving expert knowledge. Residual learning is applied to effectively address degradation problems in the deep neural network.

3. The effectiveness of the proposed model was verified on a popular bearing dataset. Compared with some existing methods, the MTF-ResNet method achieved better accuracy in bearing fault diagnosis.
4. To further demonstrate the performance of the proposed method and investigate the intrinsic mechanism of the CNN model in bearing fault diagnosis, t-SNE was used to visualize the feature maps learned by the model.

The remainder of this paper is organized as follows. Section 2 introduces the fundamentals of CNN and residual networks. In Section 3, the details of the proposed MTF-ResNet model for fault diagnosis are elaborated. Section 4 outlines experimental analysis to verify the effectiveness of the proposed model by employing a popular bearing dataset. Section 5 presents the conclusions.

2. Background and Related Work

Motivated by the concept of various cells in the visual cortex of the brain, and some cells that are exclusively responsive to the local receptive field [30], convolutional neural networks (CNNs) were first proposed by LeCun [31] for image processing. A typical CNN involves three different layers: (1) convolutional layer, (2) subsampling or pooling layers, and (3) fully connected layer. The convolutional layer comprises a number of kernels that extract features from input data. The pooling layer is the downsampling layer to reduce the trained parameters. The fully connected layer is a traditional feed-forward neural network where all neurons are connected to the activation of the previous layer. In this section, we describe CNNs and residual networks in more detail.

2.1. Convolutional Layer

The convolutional layer performs convolutional operations on local regions of the input data (or features) with the use of the convolutional kernel. Weight sharing is the most essential characteristic of the convolutional layer, since the input is traversed once by the same convolutional kernel at a set stride which can minimize the parameters and alleviate overfitting to some extent. In general, the mathematical model of the convolutional layer can be described as:

$$x_j^l = \sigma(\sum_{i \in M_j} x_j^{l-1} * k_{ij}^l + b_j^l) \tag{1}$$

where x_j^{l-1} is the input to the $(l-1)$st layer of the network; x_j^l is the output of layer l of the network; k_{ij}^l is the weight matrix of the convolution kernel; b_j^l is the bias; M_j denotes the set of input feature maps; σ represents the nonlinear activation function; $*$ represents the operation of convolution.

2.2. Pooling Layer

The main function of the pooling layer is to reduce the dimensionality of the data after convolutional operations. Average and maximal pooling are two commonly used pooling methods. The pooling layer performs a downsampling operation on the feature map, which avoids overfitting to a certain extent while retaining key features. The pooling layer transformation can be described as:

$$x_j^l = \sigma(\beta_j^l down(x_j^{l-1}) + b_j^l) \tag{2}$$

where $down(\cdot)$ represents the downsampling function, β_j^l is the multiplicative weight.

2.3. Residual Network

Traditional deep convolutional neural networks are difficult to train as the network deepens because of problems of vanishing and exploding gradients. To address the degradation problem, He et al. [29] proposed deep residual networks that are widely used in image processing. The structure of the residual networks is shown in Table 1.

Table 1. Structure of residual networks.

Layer Name	ResNet-18	ResNet-34	ResNet-50	Output Size
Conv1	7×7, 64, stride 2			112×112
Conv2_x	3×3 max pool, stride 2			56×56
	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$	
Conv3_x	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$	28×28
Conv4_x	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1028 \end{bmatrix} \times 6$	14×14
Conv5_x	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$	7×7
	Average pool, fc, softmax			1×1

Residual building blocks are the basic components of a residual network. As shown in Figure 1, a residual building block is composed of several convolutional layers, batch normalizations (BNs), ReLU activation functions, and an identity shortcut. The residual building block can be expressed as:

$$y = \mathcal{F}(\mathrm{x},\{W_i\}) + x \tag{3}$$

where x represents the input vectors of the layer and y represents the output. $\mathcal{F}(x, \{W_i\})$ denotes the residual mapping function. Take the diagram in Figure 1 for example, $\mathcal{F} = W_2\sigma(W_1x)$, where σ denotes the nonlinear activation function (ReLU).

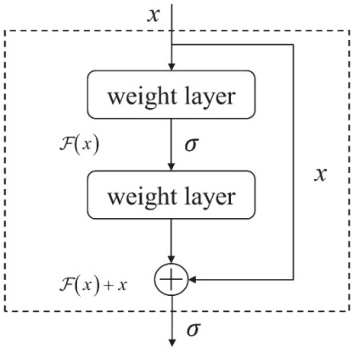


Figure 1. Residual building block.

3. Proposed Model for Fault Diagnosis

This section presents the proposed MTF-ResNet fault diagnosis method. First, data augmentation is used to increase the training data. Then, the conversion method of the vibration signals into images is presented. Lastly, the network architecture based on MTF and ResNet for rolling bearing fault diagnosis is established.

3.1. Data Augmentation

An effective technique to improve the generalization capabilities of machine-learning models is to use additional training samples. In computer vision tasks, horizontal flips,

random crops or scales, and color jitter are commonly utilized to increase the data to train the model. Data augmentation is also required in fault diagnosis for deep convolutional neural networks to achieve high classification accuracy and avoid overfitting. The data augmentation method used in this paper is overlapping samples from raw one-dimensional sequences. Augmented samples were all allocated the same fault label as that of the raw sequence, since each input sequence was obtained under a single fault state. The data augmentation process is shown in Figure 2. The specific calculation method is expressed as follows:

$$N = \frac{L - l}{s} + 1 \tag{4}$$

where L is the length of the raw signal, l is the length of a single sample, s is the shift stride, and N is the number of samples obtained through data augmentation.

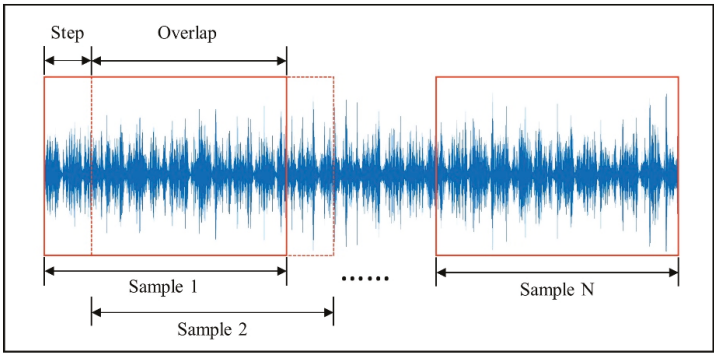


Figure 2. Process of data augmentation.

3.2. Signal-to-Image Conversion

When diagnosing and analyzing bearing faults, the accelerometer is one of the most frequently used sensors in modern research, which can directly collect the original vibration signal of the target object. Collected data from industrial processes are continuous time series, and have the characteristics of nonlinearity and nonstationary caused by high coupling in the system.

Assume a time series $X = \{x_1, x_2, \cdots, x_n\}$; the values can be quantized in Q bins, and each x_i can be allocated to a related $q_j (j \in [1, Q])$. By calculating the transitions among bins in the way of a first-order Markov chain along each time step, a matrix W of $Q \times Q$ size is obtained. $w_{i,j}$ is the probability that an element in q_j is followed by an element in q_i. After normalization by $\sum_{j=1}^{Q} w_{ij} = 1$, W is considered to be the Markov transition matrix. Since the matrix is not sensitive to the distribution of X and time steps t_i, in order to reduce the loss of information, the M_{ij} in the Markov transition field (MTF) is defined as follows:

$$M_{ij} = \begin{bmatrix} w_{ij} | x_1 \in q_i, x_1 \in q_j & \cdots & w_{ij} | x_1 \in q_i, x_n \in q_j \\ w_{ij} | x_2 \in q_i, x_1 \in q_j & \cdots & w_{ij} | x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ w_{ij} | x_n \in q_i, x_1 \in q_j & \cdots & w_{ij} | x_n \in q_i, x_n \in q_j \end{bmatrix} \tag{5}$$

The Markov transition field (MTF) then can be defined as follows:

$$M = \begin{bmatrix} M_{11} & \cdots & M_{1n} \\ M_{21} & \cdots & M_{2n} \\ \vdots & \ddots & \vdots \\ M_{n1} & \cdots & M_{nn} \end{bmatrix} \tag{6}$$

M_{ij} is the probability that an element in q_j is followed by an element in q_i. In other words, MTF incorporates temporal information on the basis of the Markov transfer matrix and actually represents the multispan transition probabilities of the time series. Such an expansion can denote not only the state transition for a single time stamp i. but also characterize state transitions over multiple time bins according to changes in the elements of MTF. $M_{ij\|i-j\|=k}$ represents the transition probability between points with a time interval k. A special case is that, when $k = 0$, main diagonal M_{ii} obtains the probability from each quantile to itself at time step i.

In the MTF matrix, the M_{ij} can be regarded to be a pixel point represented through the colormap. Red denotes a larger value, while blue denotes a smaller value. It is inappropriate to directly employ images generated by MTF as the input of CNN since the images may be too large for training the model. In order to reduce the size of the images and improve computation efficiency, blurring kernel $\left\{\frac{1}{m^2}\right\}_{m\times m}$ was adopted to average the pixels in each nonoverlapping $m \times m$ region. The transformation process of the Markov transition field is shown in Figure 3.

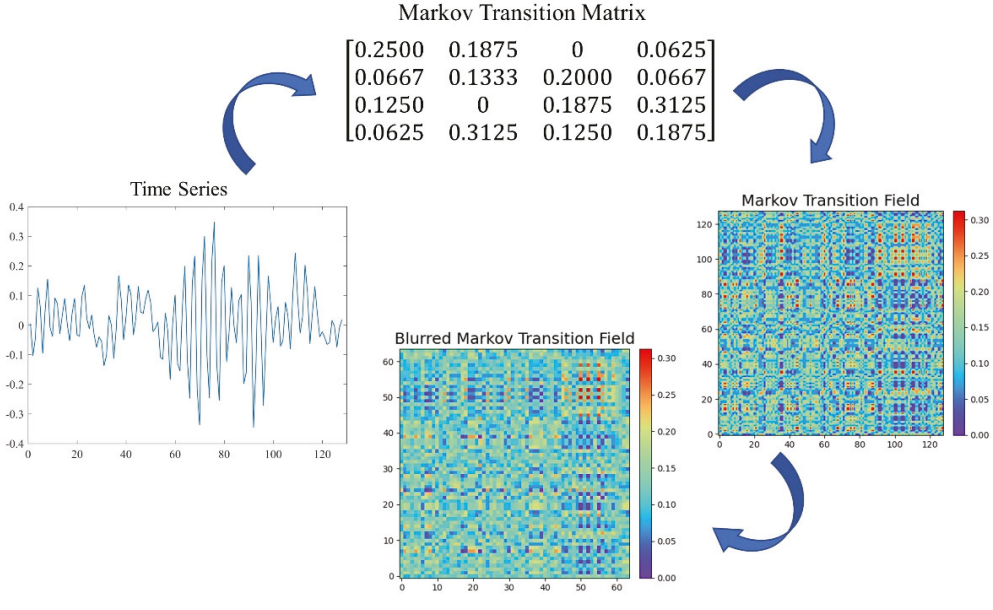


Figure 3. Transformation process of Markov transition field.

3.3. Network Architecture

Once the raw vibration signals are converted into MTF images and formed into the image dataset, a CNN model can be trained to classify these images. In this paper, we applied the ResNet-34 network to extract 2D image features. A softmax layer was employed at the end of the network to classify the rolling-bearing health condition on the basis of the learned features. The proposed MTF-ResNet model architecture is demonstrated in Figure 4. The detailed parameters of the MTF-ResNet model are presented in Table 2.

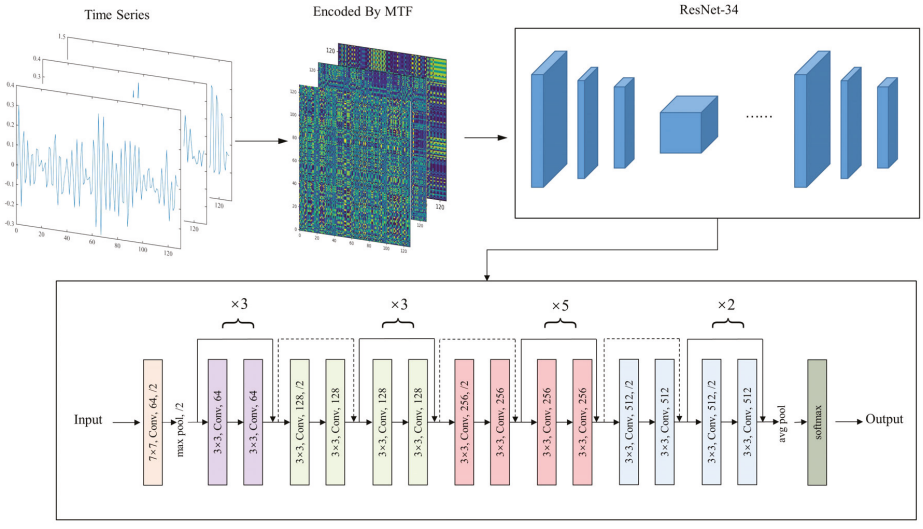


Figure 4. Architecture of the proposed MTF-ResNet model.

Table 2. Detailed parameters of the MTF-ResNet model.

Parameters	Value
Batch size	32
Optimizer	Adam
Lr	0.0001
Loss function	Category—cross-entropy

4. Experiments and Results

4.1. Data Processing

To validate the performance of the proposed MTF-ResNet, the Case Western Reserve University (CWRU) [32] bearing dataset was employed to conduct experiments. The test rig comprised an electric motor, a torque transducer/encoder, and a dynamometer, as shown in Figure 5. The bearing to be tested rotatably supports the shaft of the motor under four load conditions: 0, 1, 2 and 3 hp with motor speeds of 1772, 1750, and 1730 r/min. Different types and severity levels of bearing failures are caused by the use of electrical discharge machining (EDM), including normal condition (NC), inner-race fault (IF), outer-race fault (OF), and rolling ball fault (BF). For each fault state, three kinds of fault diameters were set: 0.007, 0.014, and 0.021 inches, respectively.

In this paper, we used raw vibration signal sample at 12 kHz from the drive end accelerometer (DE). The training data were generated from half of the raw vibration sequence by overlapping samples through a sliding window length of 2048 with a step size of 80, while the test data were generated by the same window length from the other half without data augmentation. According to the working conditions, datasets under a single working condition and variable working conditions are considered in this study. The bearing fault datasets under a single working condition are shown in Table 3; each dataset contained 6600 training samples and 250 testing samples from 10 fault types, as presented in Table 4. The composition of bearing fault data under variable working conditions is shown in Table 5.

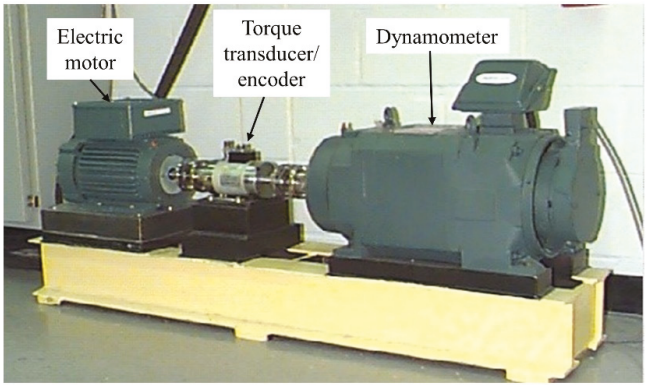


Figure 5. CWRU bearing test rig [32].

Table 3. Working conditions studied in this work.

Dataset	Motor Load (hp)	Motor Speed (r/min)
A	1	1772
B	2	1750
C	3	1730

Table 4. Composition of single working condition bearing fault data.

Fault Type	Fault Diameter (Inch)	Number of Samples	Label
BF07	0.007	660/25	0
BF14	0.014	660/25	1
BF21	0.021	660/25	2
IF07	0.007	660/25	3
IF14	0.014	660/25	4
IF21	0.021	660/25	5
NC	0	660/25	6
OF07	0.007	660/25	7
OF14	0.014	660/25	8
OF21	0.021	660/25	9

Table 5. Composition of bearing fault data under variable working conditions (Dataset D).

Fault Type	Fault Diameter (Inch)	Motor Load (hp)	Label
NC	0	0	0
IF07	0.007	1	1
BF14	0.014	2	2
OF21	0.021	3	3

All samples were then converted into MTF images. Figure 6 shows the transformation of the same signal containing 2048 data points into MTF images of different image sizes. Large MTF images generally result in an increase in computational cost and are not conducive to the training of the model. However, small MTF images can hardly contain enough useful information. On the basis of the above considerations, the size of the MTF images was determined to be 224×224.

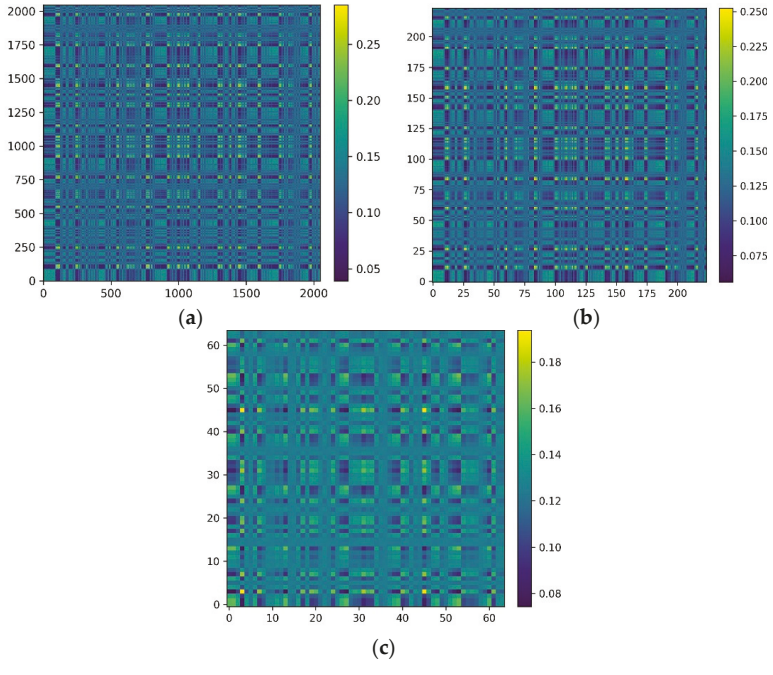


Figure 6. Transformation of the same signal containing 2048 data points into MTF images of image sizes of (**a**) 2048 × 2048, (**b**) 224 × 224, and (**c**) 64 × 64.

4.2. Data Analysis

In order to show the detailed identification effect of the model for each fault type in the test set, a confusion matrix was introduced for more accurate and comprehensive analysis of the experimental results. The vertical axis of the confusion matrix represents the true labels of the classification, and the horizontal axis demonstrates the predicted labels. The confusion matrix shows the classification results of all fault types, containing both correct and incorrect classification information. The confusion matrices of the MTF-ResNet prediction results are shown in Figure 7. In Dataset A, there was a slight error in the classification of fault types BF07 and BF21, two samples of bearing fault type BF07 were incorrectly labeled as BF21, and one sample of BF21 was identified as BF07; all other samples were correctly classified by the MTF-ResNet model. In Dataset B, the incorrect classification occurred in the identification of BF07 and OF14, two samples with the true label BF07 were incorrectly mistaken for OF14, and one sample belonging to the OF14 fault type was classified as BF07, the model achieved correct classification in all other fault types. In Dataset C, the situation was similar to that in Datasets A and B: two samples in BF07 were identified as BF21 and OF14, while one sample in each of BF21 and OF14 was misclassified as BF07. Samples of all fault types were correctly identified by the model in Dataset D. The accuracy of the model in Datasets A–D was 98.8%, 98.8%, 98.4%, and 100%, respectively. It is clear from the experimental results that almost all of the misclassifications occurred in the diagnosis of ball faults, which coincides with the findings in [32] that there are undiagnosed outer and inner race faults in the drive end bearing, probably caused by brinelling. We conducted several trials, and the average accuracy of the model in the 10- and 4-category datasets was 98.52% and 100%, respectively.

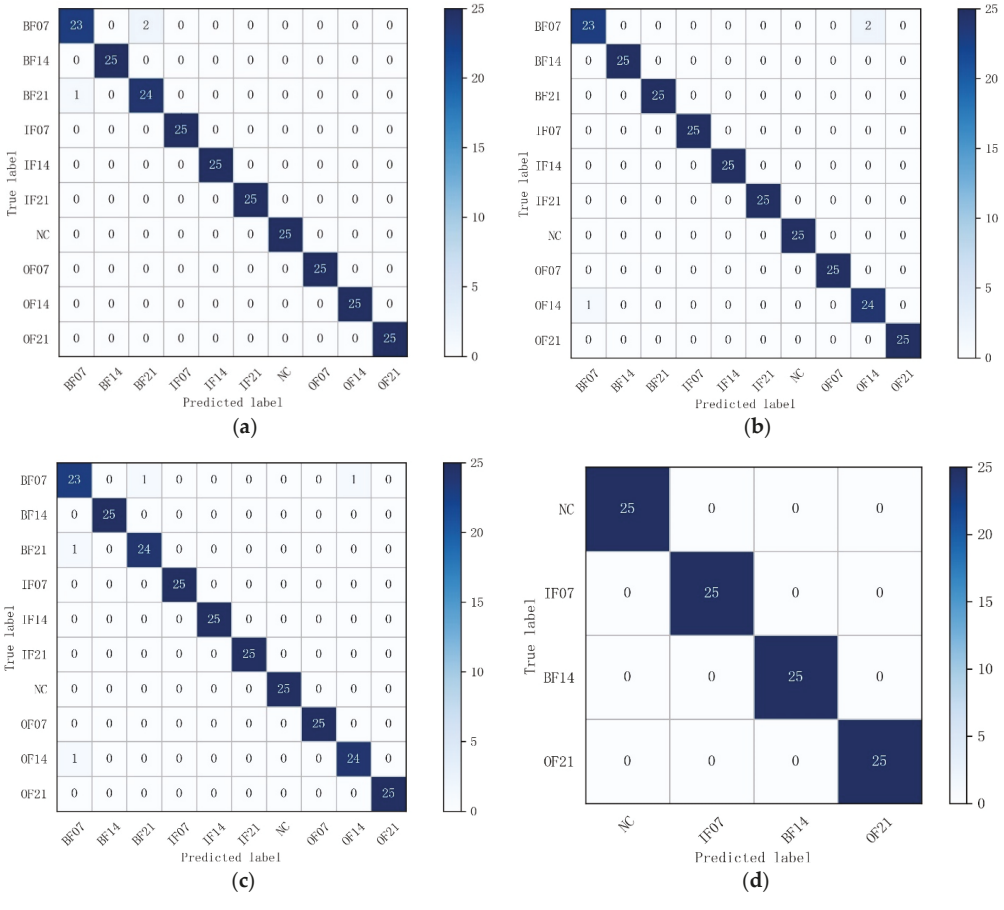


Figure 7. Confusion matrixes for each dataset: (**a**) Dataset A; (**b**) Dataset B; (**c**) Dataset C; (**d**) Dataset D.

In order to qualitatively illustrate the effectiveness of the proposed model and judge the separability of the data on the basis of the visualization of learned representation, nonlinear dimensionality reduction algorithm t-SNE was employed to project the data into a 2-dimensional space. Figure 8 shows the visualization results of the MTF-ResNet model for the 10- and 4-category datasets.

The model had powerful feature extraction and classification capability, samples of different fault types were almost perfectly separated, and samples within the same type were intuitively clustered. The results of feature visualization are consistent with the confusion matrices and demonstrate that the fault diagnosis problem can be successfully addressed by the proposed MTF-ResNet model.

To better understand the effect of convolutional layers of the model in fault diagnosis, the features extracted from the four convolutional layers are visually mapped into a two-dimensional distribution by t-SNE, as shown in Figure 9.

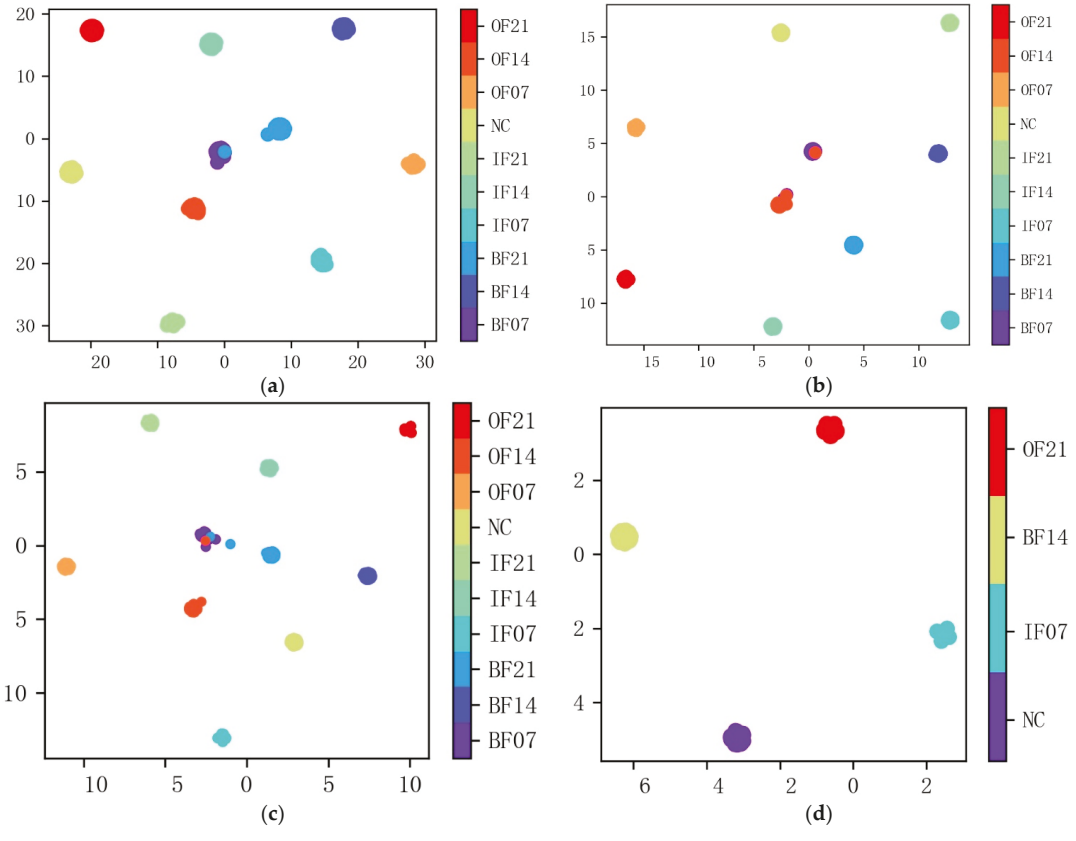


Figure 8. Feature visualization by t-SNE for each dataset: (**a**) Dataset A; (**b**) Dataset B; (**c**) Dataset C; (**d**) Dataset D.

Figure 9a shows the distribution results of the first convolutional layer, the redundancy of the vibration signal itself makes it difficult to distinguish between the different fault types. From Figure 9b, the samples of IF21, OF21 and OF07 are separated out while the rest samples of different categories are mixed. After the 23rd convolutional layer, the output sample distribution significantly changed. Most of the samples are clustered in their respective regions, but there are still some samples that are not clustered and are scattered among the adjacent categories, as shown in Figure 9c. Results of the fully connected layer are shown in Figure 9d; all samples were separated out and then clustered into their regions except for the rolling ball fault samples, which had a certain degree of misclassification.

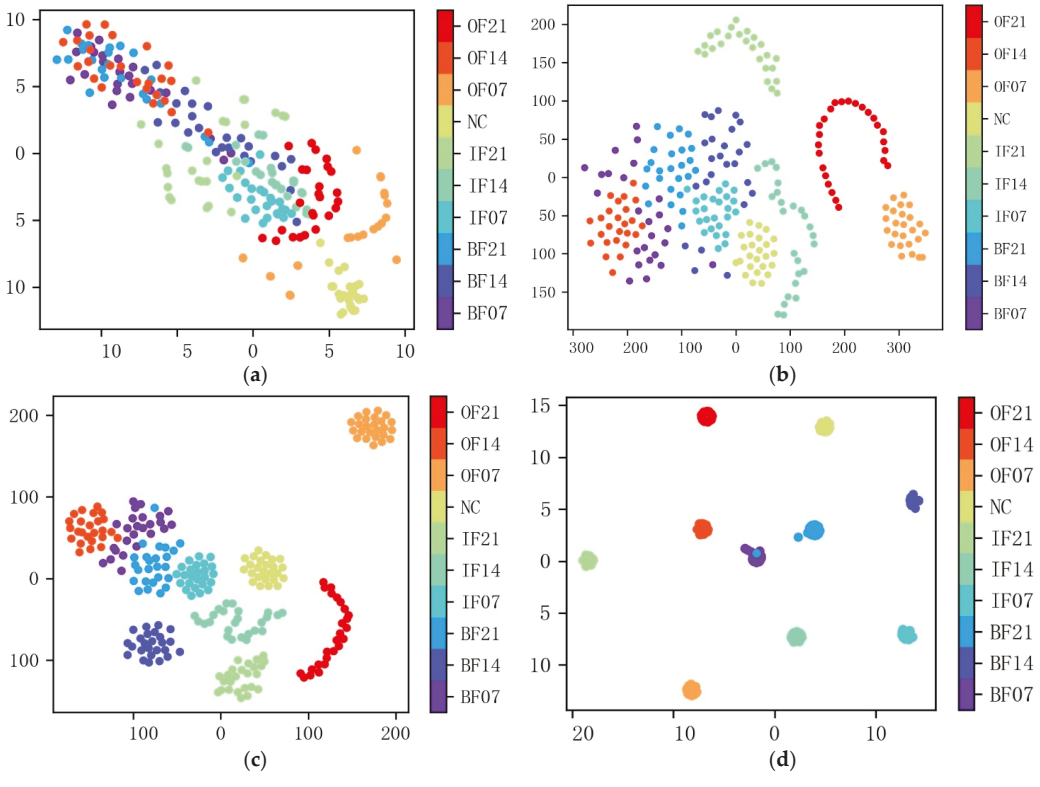


Figure 9. Feature visualization of different layers of the proposed MTF-ResNet. (**a**) First convolutional layer; (**b**) 13th convolutional layer; (**c**) 23rd convolutional layer; (**d**) fully connected layer.

4.3. Model Performance with Different Residual Network Structures

In this section, the performance of the MTF-ResNet model with different residual network structures is investigated. The same 10-category dataset was adopted, and the encoded MTF images were applied as input in ResNet-18 and ResNet-50 for feature extraction and classification. The average classification accuracy of different residual structures is shown in Table 6. It is clear that the residual networks achieved good classification accuracy of over 94% for images of bearing fault signals converted by the Markovt transition field, and the model using ResNet-34 achieved better accuracy of over 4.67% and 2.16% than that of the models using ResNet-18 and ResNet-50, respectively.

Table 6. Average classification accuracy of different residual structures.

Network	Epoch	Accuracy (%)
ResNet-18	100	94.12
ResNet-34	100	98.52
ResNet-50	100	96.44

4.4. Comparison with Other Methods

In recent years, much research has been conducted for rolling-bearing fault diagnosis problems. In order to further prove the superiority of the MTF-ResNet method proposed in this paper, we compared it with some commonly used methods. The detailed comparison results are shown in Table 7. As obtained from the experimental results, the method in [25]

could achieve 100% testing accuracy, but the model was only validated for 4-category fault classification. The proposed method could achieve an average accuracy of 98.52% for 10-category datasets and 100% for 4-category dataset. Compared with the methods in [33–36], the proposed MTF-ResNet method could identify more fault types and improve classification accuracy.

Table 7. Experimental results of different methods.

Methods	Categories	Accuracy (%)
VI-CNN [25]	4	100
STFT-CNN [33]	4	99.4
Compact 1D-CNN [34]	6	93.2
IDSCNN [35]	10	93.84
CNNEPDNN [36]	10	97.85
Proposed	4	100
	10	98.52

5. Conclusions

In this work, we proposed a novel intelligent rolling-bearing fault diagnosis method based on the Markov transition field (MTF) and residual network. Encoding one-dimensional time-series signals into two-dimensional images by Markov transition field preserves the time dependence of the raw signals and discards the prior knowledge to set parameters during the conversion. On this basis, a residual network is applied to identify the fault types through image classification. Experiments conducted on the CWRU bearing dataset indicate that MTF-ResNet achieved prominent performance on the identification of rolling bearings faults with various degrees of severity and locations, the proposed model achieves an average accuracy of 100% and 98.52% in the 4- and 10-category datasets, respectively. Compared with other intelligent bearing-fault diagnosis methods, the proposed MTF-ResNet method offers a better performance of feature extraction and classification in the fault diagnosis.

While the MTF-ResNet method can achieve good performance for fault diagnosis, it has the disadvantage of requiring a longer training period than other shallow neural network-based methods do, as the residual network in this study was trained from scratch. Deep-learning algorithms are frequently hampered by a high computational burden. In further work, the transfer-learning approach, which showed promising results in reducing training time and computational cost [37], will be considered to be employed in machinery fault diagnosis tasks. In addition, further investigations into the effectiveness of the MTF-ResNet method should be carried out a wider variety of datasets, such as gear- and rotor-fault datasets.

Author Contributions: Conceptualization, H.L. and J.Y.; methodology, J.Y. and J.K.; software, J.Y.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y., H.L. and J.K.; visualization, J.Y.; supervision, H.L. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fundamental Research Funds for the Central Universities (grant number 2017ZY46) and the National Natural Science Foundation of China (grant number 51705022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Case Western Reserve University Bearing Data https://engineering.case.edu/bearingdatacenter (accessed on 5 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, B.; He, Z.; Chen, X.; Cao, H.; Cai, G.; Zi, Y. A demodulating approach based on local mean decomposition and its applications in mechanical fault diagnosis. *Meas. Sci. Technol.* **2011**, *22*, 055704. [CrossRef]
2. Jayaswal, P.; Verma, S.N.; Wadhwani, A.K. Development of EBP-Artificial neural network expert system for rolling element bearing fault diagnosis. *J. Vib. Control* **2011**, *17*, 1131–1148. [CrossRef]
3. Yiakopoulos, C.T.; Gryllias, K.C.; Antoniadis, I.A. Rolling element bearing fault detection in industrial environments based on a K-means clustering approach. *Expert Syst. Appl.* **2011**, *38*, 2888–2911. [CrossRef]
4. Zheng, J.; Pan, H.; Cheng, J. Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines. *Mech. Syst. Signal Process.* **2017**, *85*, 746–759. [CrossRef]
5. Lei, Y.; Jia, F.; Zhou, X.; Lin, J. A deep learning-based method for machinery health monitoring with big data. *Jixie Gongcheng Xuebao J. Mech. Eng.* **2015**, *51*, 49–56. [CrossRef]
6. He, M.; He, D. Deep Learning Based Approach for Bearing Fault Diagnosis. *IEEE Trans. Ind. Appl.* **2017**, *53*, 3057–3065. [CrossRef]
7. Chen, J.; Li, Z.; Pan, J.; Chen, G.; Zi, Y.; Yuan, J.; Chen, B.; He, Z. Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review. *Mech. Syst. Signal Process* **2016**, *70–71*, 1–35. [CrossRef]
8. Guo, T.; Deng, Z. An improved EMD method based on the multi-objective optimization and its application to fault feature extraction of rolling bearing. *Appl. Acoust.* **2017**, *127*, 46–62. [CrossRef]
9. Cai, B.; Liu, H.; Xie, M. A real-time fault diagnosis methodology of complex systems using object-oriented Bayesian networks. *Mech. Syst. Signal Process.* **2016**, *80*, 31–44. [CrossRef]
10. Hajnayeb, A.; Ghasemloonia, A.; Khadem, S.E.; Moradi, M.H. Application and comparison of an ANN-based feature selection method and the genetic algorithm in gearbox fault diagnosis. *Expert Syst. Appl.* **2011**, *38*, 10205–10209. [CrossRef]
11. Saidi, L.; Ben Ali, J.; Friaiech, F. Application of higher order spectral features and support vector machines for bearing faults classification. *Isa Trans.* **2015**, *54*, 193–206. [CrossRef]
12. Tian, J.; Morillo, C.; Azarian, M.H.; Pecht, M. Motor Bearing Fault Detection Using Spectral Kurtosis-Based Feature Extraction Coupled With K-Nearest Neighbor Distance Analysis. *IEEE Trans. Ind. Electron.* **2016**, *63*, 1793–1803. [CrossRef]
13. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
14. Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H.G.; Ogata, T. Audio-visual speech recognition using deep learning. *Appl. Intell.* **2015**, *42*, 722–737. [CrossRef]
15. Abu Mallouh, A.; Qawagneh, Z.; Barkana, B.D. Utilizing CNNs and transfer learning of pre-trained models for age range classification from unconstrained face images. *Image Vis. Comput.* **2019**, *88*, 41–51. [CrossRef]
16. Shenfield, A.; Howarth, M. A Novel Deep Learning Model for the Detection and Identification of Rolling Element-Bearing Faults. *Sensors* **2020**, *20*, 5112. [CrossRef]
17. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
18. Shao, H.; Jiang, H.; Zhang, X.; Niu, M. Rolling bearing fault diagnosis using an optimization deep belief network. *Meas. Sci. Technol.* **2015**, *26*, 115002. [CrossRef]
19. Wen, L.; Gao, L.; Li, X. A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis. *IEEE Trans. Syst. Man Cybern.-Syst.* **2019**, *49*, 136–144. [CrossRef]
20. Jiang, H.; Li, X.; Shao, H.; Zhao, K. Intelligent fault diagnosis of rolling bearings using an improved deep recurrent neural network. *Meas. Sci. Technol.* **2018**, *29*, 065107. [CrossRef]
21. Hasan, M.J.; Sohaib, M.; Kim, J.-M. An Explainable AI-Based Fault Diagnosis Model for Bearings. *Sensors* **2021**, *21*, 4070. [CrossRef]
22. Chen, Z.; Mauricio, A.; Li, W.; Gryllias, K. A deep learning method for bearing fault diagnosis based on Cyclic Spectral Coherence and Convolutional Neural Networks. *Mech. Syst. Signal Process.* **2020**, *140*, 106683. [CrossRef]
23. Guo, L.; Lei, Y.; Xing, S.; Yan, T.; Li, N. Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines with Unlabeled Data. *IEEE Trans. Ind. Electron.* **2019**, *66*, 7316–7325. [CrossRef]
24. Jia, F.; Lei, Y.; Lu, N.; Xing, S. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mech. Syst. Signal Process.* **2018**, *110*, 349–367. [CrossRef]
25. Hoang, D.-T.; Kang, H.-J. Rolling element bearing fault diagnosis using convolutional neural network and vibration image. *Cogn. Syst. Res.* **2019**, *53*, 42–50. [CrossRef]
26. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method. *IEEE Trans. Ind. Electron.* **2018**, *65*, 5990–5998. [CrossRef]
27. Li, G.; Deng, C.; Wu, J.; Chen, Z.; Xu, X. Rolling Bearing Fault Diagnosis Based on Wavelet Packet Transform and Convolutional Neural Network. *Appl. Sci.-Basel* **2020**, *10*, 770. [CrossRef]
28. Wang, Z.; Oates, T. Encoding Time Series as Images for Visual Inspection and Classification Using Tiled Convolutional Neural Networks. In Proceedings of the Workshops at the Twenty-ninth AAAI Conference on Artificial Intelligence, Austin, TA, USA, 25–30 January 2015.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

30. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
31. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
32. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64–65*, 100–131. [CrossRef]
33. Pham, M.T.; Kim, J.-M.; Kim, C.H. Accurate Bearing Fault Diagnosis under Variable Shaft Speed using Convolutional Neural Networks and Vibration Spectrogram. *Appl. Sci.* **2020**, *10*, 6385. [CrossRef]
34. Eren, L.; Ince, T.; Kiranyaz, S. A Generic Intelligent Bearing Fault Diagnosis System Using Compact Adaptive 1D CNN Classifier. *J. Signal Process. Syst.* **2019**, *91*, 179–189. [CrossRef]
35. Li, S.; Liu, G.; Tang, X.; Lu, J.; Hu, J. An Ensemble Deep Convolutional Neural Network Model with Improved D-S Evidence Fusion for Bearing Fault Diagnosis. *Sensors* **2017**, *17*, 1729. [CrossRef] [PubMed]
36. Li, H. Ji Bearing Fault Diagnosis with a Feature Fusion Method Based on an Ensemble Convolutional Neural Network and Deep Neural Network. *Sensors* **2019**, *19*, 2034. [CrossRef] [PubMed]
37. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.

sensors

MDPI

Article

Rolling Bearing Fault Diagnosis Based on Successive Variational Mode Decomposition and the EP Index

Yuanjing Guo [1], Youdong Yang [1,*], Shaofei Jiang [2,*], Xiaohang Jin [2] and Yanding Wei [3]

[1] Zhijiang College, Zhejiang University of Technology, Shaoxing 312030, China; guozi@zzjc.edu.cn
[2] College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou 310023, China; xhjin@zjut.edu.cn
[3] Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, College of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China; weiyd@zju.edu.cn
* Correspondence: yydong@zjut.edu.cn (Y.Y.); jsf75@zjut.edu.cn (S.J.)

Abstract: Rolling bearing is an important part guaranteeing the normal operation of rotating machinery, which is also prone to various damages due to severe running conditions. However, it is usually difficult to extract the weak fault characteristic information from rolling bearing vibration signals and to realize a rolling bearing fault diagnosis. Hence, this paper offers a rolling bearing fault diagnosis method based on successive variational mode decomposition (SVMD) and the energy concentration and position accuracy (EP) index. Since SVMD decomposes a vibration signal of a rolling bearing into a number of modes, it is difficult to select the target mode with the ideal fault characteristic information. Comprehensively considering the energy concentration degree and frequency position accuracy of the fault characteristic component, the EP index is proposed to indicate the target mode. As the balancing parameter is crucial to the performance of SVMD and must be set properly, the line search method guided by the EP index is introduced to determine an optimal value for the balancing parameter of SVMD. The simulation and experiment results demonstrate that the proposed SVMD method is effective for rolling bearing fault diagnosis and superior to the variational mode decomposition (VMD) method.

Keywords: rolling bearing; fault diagnosis; successive variational mode decomposition; squared envelope spectrum; EP index

Citation: Guo, Y.; Yang, Y.; Jiang, S.; Jin, X.; Wei, Y. Rolling Bearing Fault Diagnosis Based on Successive Variational Mode Decomposition and the EP Index. *Sensors* **2022**, *22*, 3889. https://doi.org/10.3390/s22103889

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 24 April 2022
Accepted: 13 May 2022
Published: 20 May 2022

1. Introduction

Rolling bearing is an important part widely used in rotating machinery, such as wind turbines [1], high-speed railways [2], helicopters [3] and electric vehicles [4]. As rolling bearing always operates under harsh working conditions, ranging from high speed and alternating speed to heavy load and alternating load, its inner race, outer race, and balls are prone to suffer from various kinds of damages, including fatigue pitting, wear, spalling, cracking, etc. Hence, it is of great significance to take effective technical measures for rolling bearing health condition monitoring and fault diagnosis to guarantee the reliable operation and long-term economic benefits of rotating machinery. When a localized defect occurs in the inner raceway, outer raceway, or balls of a rolling bearing, a series of impact events are excited because the damaged rolling contact surface lacks smooth support [5]. Specifically, for a rolling bearing under steady working conditions, these impact events are presented as periodic impulse features in the vibration signal. Different impulse feature frequencies are closely related to the fault characteristic frequencies (FCFs) of different components of a rolling bearing. Therefore, it comes very naturally to extract the fault characteristic frequency (FCF) from sampled vibration signal and then diagnose a rolling bearing fault. However, as a rolling bearing is usually installed inside the machinery together with other rotating parts, the impulsive vibration features excited by fault damage will attenuate along the transfer path from the source to the location of the vibration sensor. Based on this, as

well as taking into consideration background noise and interference components caused by other parts, the weak fault features' extraction from a vibration signal for rolling bearing fault diagnosis is usually a challenging task and has attracted considerable attention. Thus, many useful methods have been developed.

One of the most effective and practical methods is signal decomposition, including wavelet packet decomposition [6], empirical mode decomposition (EMD) [7,8], local mean decomposition (LMD) [9,10], empirical wavelet transform (EWT) [11–14], variational mode decomposition (VMD) [15–18], and so on. Signal decomposition can extract the useful component containing fault characteristic information and thus achieve the purpose of removing noise and interference components. Among the aforementioned signal decomposition methods, VMD decomposes a signal into an ensemble of band-limited sub-signals called modes [19], and it currently receives extensive study and application due to its complete mathematical principles and ability to avoid the shortcomings of sensitivity to noise, end effects, and mode mixing, which are inherent in EMD and LMD [20–23]. Nevertheless, there are two critical parameters affecting the performance of VMD that need to be pre-set properly for VMD implementation: the balancing parameter and number of modes. Although these two parameters can be directly pre-set by experience or experiments, the method has the drawback of blindness and is difficult to obtain excellent performance of VMD. Consequently, researchers usually utilize some intelligent optimization algorithms to determine the values of the two parameters, such as the genetic algorithm [24,25], particle swarm optimization [16,26], differential search algorithm [27], Archimedes optimization algorithm [28], grey wolf optimization [29,30], whale optimization algorithm [31], cuckoo search algorithm [32], sparrow search algorithm [33], and so on.

Although these intelligent optimization algorithms have achieved successful applications for the determination of VMD parameters, there are still some drawbacks that cannot be ignored. The implementation of an intelligent optimization algorithm usually requires a lot of initializations and iterative calculations, which is a highly time-consuming process. In addition, the intelligent optimization algorithm can easily fall into the local optimal value or even be difficult for convergence. Therefore, finding the optimal values for the two parameters of VMD using an intelligent optimization algorithm still needs further study.

As an improvement of VMD, a novel signal decomposition method known as variational mode extraction (VME) was proposed by Nazari and Sakhaei [34]. VME is homologous with VMD but decomposes a signal into two modes, the desired mode and residual signal, which avoids the trouble of determining the modes number associated with VMD. However, in the application of a rolling bearing fault diagnosis, it is often difficult to determine the initial center frequency and also not easy to optimize the balancing parameter for VME [35]. Based on this, Nazari and Sakhaei proposed an efficient and fast adaptive signal decomposition method named Successive-VMD (SVMD) [36]. In essence, the implementation of SVMD is done by successively applying VME on a signal. SVMD not only avoids the need to know the number of modes and has lower computational complexity in contrast to VMD, but also skirts the problem of initial center frequency determination in VME. Nevertheless, SVMD is confounded by the basic trouble of difficulty optimizing the balancing parameter. Different balancing parameter values would likely lead to different numbers of modes. Moreover, how to select the useful mode containing fault characteristic information from the multiple modes obtained by SVMD also remains an important issue needing further study.

Based on the above analysis, this paper proposes a rolling bearing fault diagnosis method based on SVMD, in which the target mode containing ideal fault characteristic information is selected from the modes using a novel index, named the energy concentration and position accuracy (EP) index. Accordingly, the line search method is adopted to achieve the globally optimal value for the balancing parameter of SVMD. The subsequent sections of this paper are organized as follows. The SVMD algorithm is briefly introduced in Section 2. Section 3 explains the principle of the proposed EP index. In Section 4, the line

search method for the optimal value of the balancing parameter and the corresponding SVMD method for rolling bearing fault diagnosis are summarized. The simulation signal analysis using the SVMD method and the EP index is described in Section 5. Three datasets associated with three kinds of rolling bearing faults are used to verify the effectiveness of the proposed method in Section 6. Concluding remarks are presented in Section 7.

2. Successive Variational Mode Decomposition

Essentially, the SVMD of a signal $x(t)$ requires successively performing VME on the signal $x(t)$ until all modes are extracted, or the reconstruction error, defined as the error between the original signal $x(t)$ and sum of the modes, is less than a given threshold [36]. In order to achieve the SVMD method, the original signal $x(t)$ is first assumed to be decomposed into two signals, i.e., the Lth mode $u_L(t)$ and residual signal $x_\mathrm{r}(t)$:

$$x(t) = u_L(t) + x_\mathrm{r}(t) \tag{1}$$

where the residual signal $x_\mathrm{r}(t)$ is also composed of two parts, i.e., the sum of the previously obtained modes and the unprocessed part $x_\mathrm{u}(t)$ of the original signal $x(t)$:

$$x_\mathrm{r}(t) = \sum_{i=1}^{L-1} u_i(t) + x_\mathrm{u}(t) \tag{2}$$

The SVMD method for the Lth mode $u_L(t)$ extraction is established based on four criteria, which are briefly described as follows [36]: (1) Each mode should be compact around its center frequency; (2) the spectral overlap between $u_L(t)$ and $x_\mathrm{r}(t)$ should be minimized; (3) the energy of $u_L(t)$ at frequencies around the center frequencies of the previously obtained modes should also be minimized; (4) the original signal $x(t)$ should be completely reconstructed from the L modes and $x_\mathrm{u}(t)$. Hence, when the $L-1$ modes are known, the task of the Lth mode extraction can be transformed into a constrained minimization problem, as follows:

$$\begin{aligned} &\min_{u_L,\omega_L,x_\mathrm{r}} \left\{ \alpha \left\| \partial_t \left[\left(\delta(t) + \tfrac{\mathrm{j}}{\pi t} \right) * u_L(t) \right] \mathrm{e}^{-\mathrm{j}\omega_L t} \right\|_2^2 + \left\| \beta_L(t) * x_\mathrm{r}(t) \right\|_2^2 + \sum_{i=1}^{L-1} \left\| \beta_i(t) * u_L(t) \right\|_2^2 \right\} \\ &\text{subject to } : u_L(t) + x_\mathrm{r}(t) = x(t) \end{aligned} \tag{3}$$

where ω_L is the center frequency of the Lth mode, α is a parameter for balancing, ∂t denotes the partial derivative with time t, $\delta(t)$ is the Dirac function, and * is the convolution operator. $\beta(t)$ is the impulse response of the filter $\hat{\beta}(\omega)$ used to filter the frequencies in $x_\mathrm{r}(t)$ overlapping with $u_L(t)$ to satisfy criterion (2). The filter $\hat{\beta}(\omega)$ can be expressed as:

$$\hat{\beta}(\omega) = \frac{1}{\alpha(\omega - \omega_L)^2} \tag{4}$$

where $\beta_i(t)$ is the impulse response of the filter $\hat{\beta}_i(\omega)$ used to filter the frequencies in $u_i(t)$ overlapping with $u_L(t)$ to satisfy criterion (3). The filter $\hat{\beta}_i(\omega)$ can be expressed as:

$$\hat{\beta}_i(\omega) = \frac{1}{\alpha(\omega - \omega_i)^2}, i = 1, 2, \cdots, L-1 \tag{5}$$

To convert the constrained minimization problem described in Equation (3) into an unconstrained optimization problem, the quadratic penalty term and Lagrangian multiplier λ are jointly introduced to establish the augmented Lagrangian function, as follows:

$$\begin{aligned} L(u_L, \omega_L, \lambda) = \; & \alpha \left\| \partial_t \left[\left(\delta(t) + \tfrac{\mathrm{j}}{\pi t} \right) * u_L(t) \right] \mathrm{e}^{-\mathrm{j}\omega_L t} \right\|_2^2 + \left\| \beta_L(t) * x_\mathrm{r}(t) \right\|_2^2 + \sum_{i=1}^{L-1} \left\| \beta_i(t) * u_L(t) \right\|_2^2 \\ & + \left\| x(t) - \left(u_L(t) + \sum_{i=1}^{L-1} u_i(t) + x_\mathrm{u}(t) \right) \right\|_2^2 + \left\langle \lambda(t), x(t) - \left(u_L(t) + \sum_{i=1}^{L-1} u_i(t) + x_\mathrm{u}(t) \right) \right\rangle \end{aligned} \tag{6}$$

According to the Parseval's theorem, Equation (6) can be converted to the frequency domain form and be rewritten as:

$$\begin{aligned} L(u_L,\omega_L,\lambda) = \quad & \alpha\|\mathrm{j}(\omega-\omega_L)[(1+\mathrm{sgn}(\omega))\cdot\hat{u}_L(\omega)]\|_2^2+\|\hat{\beta}_L(\omega)\cdot\left(\hat{x}_u(\omega)+\sum_{i=1}^{L-1}\hat{u}_i(\omega)\right)\|_2^2+\sum_{i=1}^{L-1}\|\hat{\beta}_i(\omega)\cdot\hat{u}_L(\omega)\|_2^2 \\ & +\|\hat{x}(\omega)-\left(\hat{u}_L(\omega)+\hat{x}_u(\omega)+\sum_{i=1}^{L-1}\hat{u}_i(\omega)\right)\|_2^2+\left\langle\hat{\lambda}(\omega),\hat{x}(\omega)-\left(\hat{u}_L(\omega)+\hat{x}_u(\omega)+\sum_{i=1}^{L-1}\hat{u}_i(\omega)\right)\right\rangle \end{aligned} \tag{7}$$

As in the VMD and VME methods, the alternate direction method of multipliers (ADMM) algorithm is also used to iteratively solve the above minimization problem, and the specific solution process can be seen in Reference [36]. The final iteratively updating equations of $\hat{u}_L(\omega)$, ω_L, and $\hat{\lambda}(\omega)$ are given as follows:

$$\hat{u}_L^{n+1}(\omega)=\frac{\hat{x}(\omega)+\alpha^2(\omega-\omega_L^n)^4\cdot\hat{u}_L^n(\omega)+\frac{\hat{\lambda}(\omega)}{2}}{\left[1+\alpha^2(\omega-\omega_L^n)^4\right]\cdot\left[1+2\alpha(\omega-\omega_L^n)^2+\sum_{i=1}^{L-1}\frac{1}{\alpha^2(\omega-\omega_i^n)^4}\right]} \tag{8}$$

$$\omega_L^{n+1}=\frac{\int_0^\infty\omega\left|\hat{u}_L^{n+1}(\omega)\right|^2\mathrm{d}\omega}{\int_0^\infty\left|\hat{u}_L^{n+1}(\omega)\right|^2\mathrm{d}\omega} \tag{9}$$

$$\hat{\lambda}^{n+1}=\hat{\lambda}^n+\tau\left[\hat{x}(\omega)-\left(\hat{u}_L^{n+1}(\omega)+\frac{\alpha^2(\omega-\omega_L^{n+1})^4\left(\hat{x}(\omega)-\hat{u}_L^{n+1}(\omega)-\sum_{i=1}^{L-1}\hat{u}_i(\omega)\right)-\sum_{i=1}^{L-1}\hat{u}_i(\omega)}{1+\alpha^2(\omega-\omega_L^{n+1})^4}+\sum_{i=1}^{L-1}\hat{u}_i^{n+1}(\omega)\right)\right] \tag{10}$$

where $\hat{x}(\omega)$ represents the Fourier transform of the original signal $x(t)$, $\hat{u}_L^n(\omega)$ represents the Fourier transform of the Lth mode $u_L^n(t)$ in the nth iteration with the center frequency ω_L^n, n is the number of iterations, and τ is the iteration step length. Accordingly, the complete algorithm for SVMD is described in Algorithm 1 [36].

Algorithm 1. SVMD

Input $x(t)$
Set α, ε_1, ε_2 and σ^2
Initialize, $L\leftarrow 0$
repeat
  $L\leftarrow L+1$
  Initialize $\hat{u}_L^1$, $\hat{\lambda}^1$, ω_L^1, $L\leftarrow 0$
  repeat
    $n\leftarrow n+1$
      (1) Update $\hat{u}_L$ according to Equation (8) for all $\omega\geq 0$
      (2) Update ω_L according to Equation (9)
      (3) Update $\hat{\lambda}$ according to Equation (10) using Dual Ascent method for all $\omega\geq 0$
  until convergence: $\frac{\|\hat{u}_L^{n+1}(\omega)-\hat{u}_L^n(\omega)\|_2^2}{\|\hat{u}_L^n(\omega)\|_2^2}<\varepsilon_1$
until convergence: $\left|\sigma^2-\frac{1}{T}\|x(t)-\sum_{i=1}^{L}u(t)\|_2^2\right|/\sigma^2<\varepsilon_2$

Based on the process of Algorithm 1, SVMD can be considered as the solution of K optimization problems or the solutions of K one-dimensional optimization problems at each frequency, and thus has a lower computational complexity than VMD, which is a solution of the K-dimensional optimization problem [36]. Such a superiority of SVMD over VMD was verified in Reference [36]. During the implementation of SVMD, the update parameter τ is often set as zero to accelerate the algorithm convergence. The values of the convergence tolerance, ε_1 and ε_2, can be set to small positive values in accordance with different requirements. σ^2 is an approximate value of the additive white noise power in the

original signal $x(t)$, which can be estimated using some filters, such as the Savitzky–Golay filter. The most important parameter in the SVMD algorithm is the balancing parameter α. A small α value may cause the mode mixing problems [36]. For a rolling bearing fault feature extraction, mode mixing means the fault characteristic mode may be seriously interfered by other components or noise. However, if the α value is too high, a lot of modes may be generated, most of which are noise or interference components, increasing the difficulty of the useful target mode selection, and the algorithm convergence may be affected. Therefore, the α value determination for SVMD is a very important but challenging task, as the proper α value usually varies in a large range for different signals. In addition, as there are a number of modes obtained by SVMD, it is often difficult to select the target mode containing the ideal fault characteristic information for a rolling bearing fault diagnosis.

In view of the existing shortcomings in SVMD, we propose a novel index named as the EP index to evaluate the modes obtained by SVMD and accordingly take the line search method to achieve the optimal value of the balancing parameter. Based on these studies, we propose a rolling bearing fault diagnosis method and use a simulated vibration signal of a faulty rolling bearing and three experimental vibration datasets from a rolling bearing testbed to evaluate the performance of the proposed method.

3. The EP Index

In this section, we propose the EP index to evaluate the modes obtained by SVMD and indicate the target mode among the modes. The EP index is based on the squared envelope spectrum (SES) analysis of each mode, and its principle is described in detail as follows.

3.1. Squared Envelope Spectrum

The purpose of bearing vibration signal processing is to extract the fault features, the most important of which is the FCF. An efficient and direct method for FCF extraction from a bearing vibration signal is the SES analysis. Given a vibration signal $x(t)$ of a rolling bearing, its SES calculation mainly includes three steps [37].

Step 1: Filter $x(t)$ around a resonance frequency to remove noise and highlight the structural natural vibration characteristics caused by the impact excitation of bearing damage, and the band-pass filtered signal is expressed as $x_{\mathrm{f}}(t)$.

Step 2: Calculate the squared absolute value of the analytic signal of $x_{\mathrm{f}}(t)$ to obtain the squared envelope (SE) signal:

$$SE(t) = |x_{\mathrm{f}}(t) + \mathrm{j} \cdot \mathbf{H}[x_{\mathrm{f}}(t)]|^2 \tag{11}$$

where $\mathbf{H}(\cdot)$ denotes the Hilbert transform.

Step 3: Calculate the squared absolute value of the Fourier transform (FT) of the squared envelope signal $SE(t)$ to obtain the SES:

$$SES(f) = |FT[SE(t)]|^2 \tag{12}$$

where $FT(\cdot)$ denotes the Fourier transform.

3.2. The EP Index

After a vibration signal of rolling bearing is processed by SVMD, how to evaluate the performance of each extracted mode is of great significance, which has important influence on the value determination of the key parameter α in SVMD and the useful target mode selection. As a rolling bearing mainly consists of an inner race, outer race, and balls, which are prone to damage, we focus on the three corresponding FCFs, i.e., the inner race FCF (f_{ir}), the outer race FCF (f_{or}), and the ball FCF (f_{ba}), of which the theoretical values of a given rolling bearing can be calculated directly. For a mode extracted by SVMD, if it is the expected target mode that contains complete and pure fault characteristics, its SES should satisfy two conditions, described as follows.

(1) Energy concentration: The energy of the SES should concentrate around one of the rolling bearing FCFs as much as possible, and the energy at other frequencies should be as little as possible. Hence, in the SES of the target mode, the amplitude corresponding to the possible FCF should be maximal, whereas the amplitudes at other frequencies should be very low. Accordingly, we propose an index to evaluate the energy concentration of the possible FCF component in the SES of a mode. This index is named the energy concentration (EC) index and calculated with the following two steps.

Step 1: Normalize the SES amplitudes using the following equation:

$$NS(n) = \frac{SES(n)}{\max[SES(n)]} \tag{13}$$

where $SES(n)$ (n = 1, 2, ... , N) is the discrete form of $SES(f)$, and N is the number of frequency points.

Step 2: Sort the amplitudes of $NS(n)$ in descending order to obtain the amplitude sequence $SNS(n)$ and calculate the average value of the differences between the first amplitude and the following K amplitudes to obtain the EC index:

$$EC = \frac{1}{K}\sum_{k=1}^{K}[SNS(1) - SNS(1+k)] \tag{14}$$

where, in this paper, the value of K is uniformly set as 10. According to the calculation process, $0 \leq EC \leq 1$ and the EC index of the target mode should be maximal.

(2) Position accuracy: In the SES of the target mode, the frequency corresponding to the maximum energy should be one of the rolling bearing FCFs. It also means that the deviation between the frequency corresponding to the maximum amplitude of the SES and one of the rolling bearing FCFs should be minimized. Consequently, we also propose an index to evaluate the position accuracy of the maximum amplitude of the SES. This index is named the position accuracy (PA) index and calculated using the following two steps.

Step 1: Find the frequency value corresponding to the maximum amplitude of the SES, which is expressed as f_{ma}.

Step 2: Calculate the PA index as follows:

$$PA = \left(\prod_{m=1}^{M}|f_{\mathrm{ma}} - f_{\mathrm{m}}|\right)^{1/M} \tag{15}$$

where, in this paper, f_m (m = 1, 2, ... , M) are the FCFs of the inner race, outer race, and ball of the rolling bearing, and thus, $M = 3, f_1 = f_{\mathrm{ir}}, f_2 = f_{\mathrm{or}}, f_3 = f_{\mathrm{ba}}$. When one of the three components of the rolling bearing fails, its FCF extracted from the vibration signal will make the PA index close to or equal to zero.

Combining the EC index with the PA index, we propose a new comprehensive index to evaluate the rolling bearing-related fault characteristic information in the extracted modes by SVMD. This index is named the energy concentration and position accuracy (EP) index and defined as follows:

$$EP = \frac{1}{EC^p} + \beta \cdot PA \tag{16}$$

where p is the adjustment coefficient and β is the balancing coefficient. Actually, the value of the PA index is several orders of magnitude larger than the value of the EC index and, thus, in order to compensate for this numerical gap, β can be calculated directly as follows:

$$\beta = \frac{EC}{EC + PA} \tag{17}$$

With regard to the inner race or outer race damage of the rolling bearing, the corresponding FCF represented in the SES of the target mode is usually lightly interfered by

other frequency components and can achieve a stable value close to the theoretical value. There is no need to make too many adjustments between the EC and PA index, hence, $p = 1$. However, due to the effect of random slippage of the balls, the FCF of balls is difficult to maintain as a stable value and may deviate greatly from the theoretical value. In this situation, there should be a larger weight for the PA index. In other words, the weight for the EC index should be shrunk, and p should be taken a positive value less than 1, such as $p = 0.5$. For the expected target mode extracted from a rolling bearing vibration signal, it should contain complete and sufficient fault characteristic information, and the EP index value of its SES should be minimized. Based on the above analysis, the EP index can be used to optimize the balancing parameter α for SVMD and to select the target mode from the results of SVMD.

4. The Rolling Bearing Fault Diagnosis Method Based on SVMD and the EP Index

Considering the SVMD algorithm principle and the vibration signal characteristics of a faulty rolling bearing, this paper proposes a rolling bearing fault diagnosis method based on SVMD combined with the EP index. In the original SVMD method, there is actually only one key parameter, i.e., the balancing parameter α. Therefore, for simplicity, the line search method is directly used to obtain a globally optimal value for the balancing parameter α in a given range. Since the SVMD algorithm is not very sensitive to α as long as the value of α varies within a narrow range, the step length of α variation can be taken to be a relatively large value for a fast search. The flow chart of the proposed rolling bearing fault diagnosis method is shown in Figure 1. The main steps are described as follows.

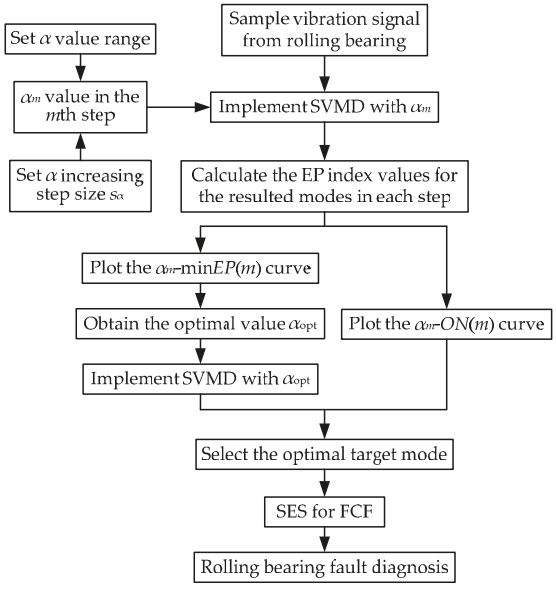


Figure 1. Process flow of rolling bearing fault diagnosis based on SVMD and the EP index.

Step 1: The value range of α is set as $[\alpha_{min}, \alpha_{max}]$, and its increasing step length is set as s_α. Then, α increases step by step from α_{min}, and in the mth step, the α value is expressed as $\alpha_m = \alpha_{min} + (m - 1) \cdot s_\alpha$.

Step 2: The SVMD algorithm is implemented with the corresponding parameter balancing α_m in each step for the rolling bearing vibration signal to obtain a series of modes. The value of the EP index for each mode is calculated, and then the mode with the minimum EP value expressed as min$EP(m)$ is selected as the target mode, whose order number is the nth mode expressed as $ON(m)$.

Step 3: After α increases to $\alpha_{\max}$ and the relevant calculation of the last step is completed, the α_m–min$EP(m)$ and α_m–$ON(m)$ curves are plotted. In the α_m–min$EP(m)$ curve, the α value corresponding to the minimum min$EP(m)$ value is selected as the globally optimal α value and expressed as α_{opt}.

Step 4: The SVMD algorithm is implemented with the balancing parameter value of α_{opt} for the rolling bearing vibration signal to obtain a series of modes, among which the optimal target mode is selected according to the order number at α_{opt} in the α_m–$ON(m)$ curve.

Step 5: SES analysis is performed for the optimal target mode to extract the FCF and diagnose the rolling bearing fault.

5. Simulation Analysis

In this section, a simulated vibration signal of a faulty rolling bearing is constructed to evaluate the efficiency of the proposed method. Considering a rolling bearing running with constant speed and assuming its inner race, outer race, or rollers have local damage, the excited vibration signal can be modeled as a series of periodic transient impulse features [38,39], and the vibration acceleration signal $x(t)$ measured from the rolling bearing can be modeled as Equations (18)–(20):

$$x(t) = s(t) + n(t) \tag{18}$$

$$s(t) = \sum_{m=1}^{M} A_m e^{-\zeta(t - mT_p - \sum_{i=1}^{m}\tau_i)} cos\left(2\pi f_n(t - mT_p - \sum_{i=1}^{m}\tau_i)\right) \cdot u(t - mT_p - \sum_{i=1}^{m}\tau_i) \tag{19}$$

$$A_m = 1 + a_m \cdot \cos(2\pi f_r t) \tag{20}$$

where $s(t)$ is an ideal impulsive vibration signal with no noise; $n(t)$ is white Gaussian noise; M is the number of the fault impulses induced by the local damage; A_m is the amplitude of the mth fault impulse; a_m is the amplitude modulation coefficient, where $0 < a_m < 1$; f_r is the rotating frequency of the bearing; ζ is the structural damping coefficient; T_p is the time period between two consecutive fault impulses, and $T_p = 1/f_c$, in which f_c represents the FCF of inner race, outer race, or balls; τ_i (i = 1, 2, . . . , M) represents the effect of random slippage of the balls and can be assumed to be a zero mean, uniformly distributed random sequence with a standard deviation of 0.01 T_p~0.02 T_p; ω_r is the excited resonance frequency; and $u(t)$ represents the unit step function.

The vibration signal of a faulty rolling bearing can be generated by setting the appropriate values for the relevant parameters. The parameters are set as ζ = 700 N.s/m, f_c = 120 Hz, T_p = 0.0083 s, and $\omega_r = 8000\pi$ rad/s. The amplitude sequences a_m (m = 1, 2, . . . , M) are obtained from a normal distribution with a mean of 0.5 and a standard deviation of 0.3. The standard deviation of the random sequences τ_i (i = 1, 2, . . . , M) is set as 0.015 T_p. The sampling frequency f_s is set as 16,000 Hz, and the number of sampling points is 3000. The vibration signal $s(t)$ with ideal fault impulse features is shown in Figure 2. Then, $s(t)$ is mixed with white Gaussian noise $n(t)$ to achieve a simulated vibration signal $x(t)$ with a signal-to-noise ratio (SNR) of −13 dB. The simulation vibration signal $x(t)$ is shown in Figure 3a. The SNR of $x(t)$ is so low that the fault-related impulsive features are completely overwhelmed by the noise and almost impossible to identify. The FCF is also unable to be extracted by SES, as shown in Figure 3b, in which the amplitudes of SES are normalized using the division-by-maximum method, and thus the ordinate represents the normalized amplitude, abbreviated as Norm. Amp.

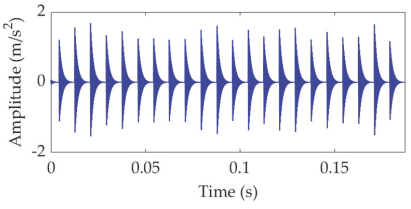
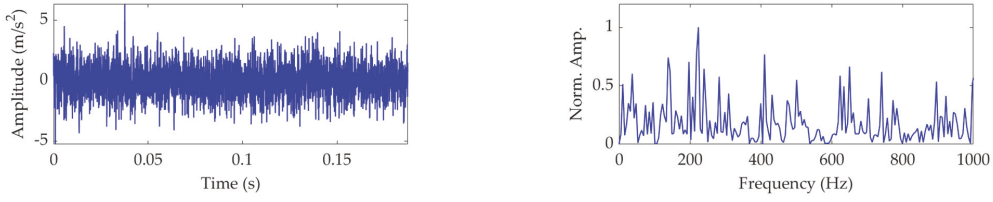


Figure 2. Vibration signal $s(t)$ with fault-related impulsive features.



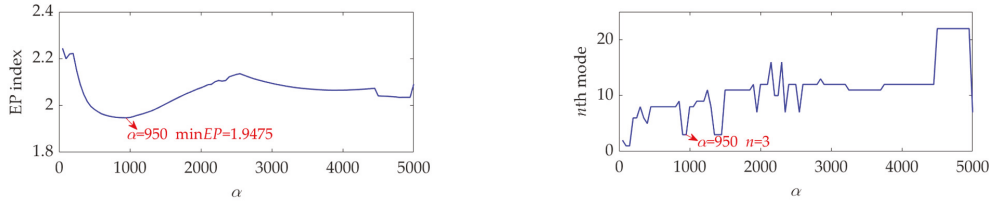
(**a**) (**b**)

Figure 3. The simulated vibration signal $x(t)$: (**a**) time−domain waveform and (**b**) SES.

To apply the SVMD algorithm to process the simulated vibration signal $x(t)$, we first need to find a relatively optimal value for the balancing parameter α of SVMD. Subsequently, we let α increase gradually from a minimum value of 50 to a maximum value of 5000 at a step size of 50. In each step of α increase, $x(t)$ is decomposed using the SVMD algorithm with the corresponding α value to obtain a series of modes, and the EP index value is calculated for each mode. One thing to note is that, in this simulated signal, only one FCF is involved, i.e., f_c = 120 Hz. The calculation of the PA index shown in Equation (15) should be accordingly modified as follows:

$$PA = |f_{ma} - f_c| \tag{21}$$

Then, the mode with the minimum EP index value is selected as the target mode in the current step. After α increases to the maximum value and the last target mode is achieved, we can draw the relationship curve between the α value and the EP index value of the target mode in each step of α increase, which is shown in Figure 4a. At the same time, we can also draw the relationship curve between the α value and the order number of the target mode among the decomposed modes in each step, which is shown in Figure 4b.



(**a**) (**b**)

Figure 4. Relationship between the α value and (**a**) the EP index value and (**b**) order number of the target mode in each step of α increase.

Figure 4a shows that when the α value is 950, the EP index of the obtained target mode acquires a global minimum value, and thus the optimal value of α is achieved, expressed as α_{opt} = 950. Meanwhile, Figure 4b shows that, when the α value is 950, the target mode is the

third mode. What we need to highlight is that, as the α value varies, both the total number of modes obtained by SVMD and the order number of the target mode are very likely to change accordingly. When the α value is large, there may be a large number of modes obtained due to narrow bandwidth, and it is always difficult to select the target mode. If the α value is set improperly, the selected target mode may not be a useful mode that can be used for the rolling bearing fault diagnosis. The existences of these problems reflect the necessity of this study, and the solutions of these problems represent the significance of this paper.

Now, we set the balancing parameter α of SVMD as α_{opt} = 950, and the decomposition results along with the corresponding squared envelope spectra of all modes are shown in Figure 5. The EP index values of all modes are shown in Figure 6. It can be seen that the total number of the modes is 5 and the third mode achieves the minimum value for the EP index, which is consistent with the result shown in Figure 4b. Therefore, the optimal target mode is the third mode, with details shown in Figure 7a, and its specific SES shown in Figure 7b. The extracted FCF is f = 122.667 Hz, which is basically consistent with the theoretical FCF of 120 Hz. The difference between the two values is caused by the effect of random slippage of the balls, considered in the simulated signal model, and the heavy noise. The process and results of the simulation analysis validate that the proposed EP index can effectively indicate the target mode from the results of SVMD, and the proposed method can successfully extract the simulated fault feature from the vibration signal with a low SNR, and thus be used for rolling bearing diagnosis.

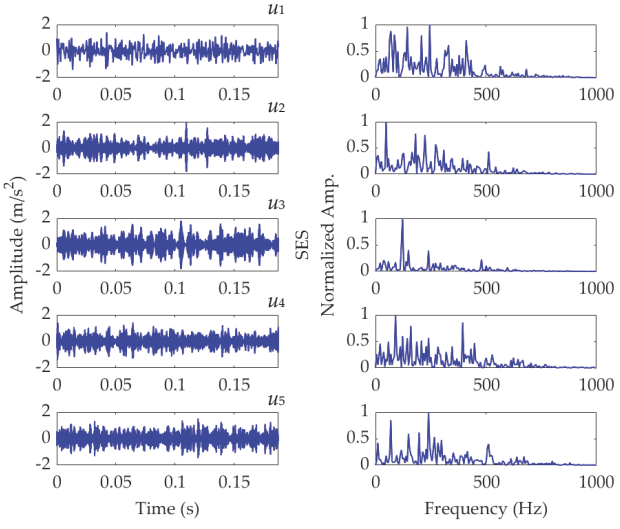


Figure 5. Results of the simulated vibration signal $x(t)$ obtained using SVMD with α_{opt} = 950.

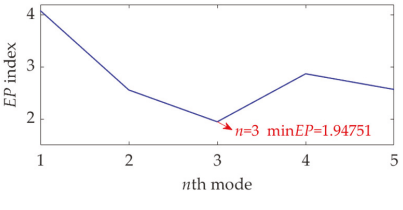


Figure 6. The EP index values of all modes.

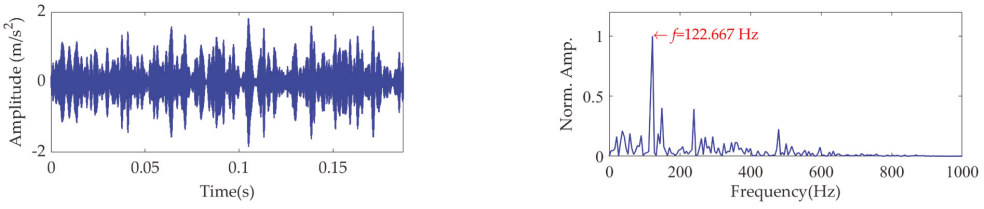


Figure 7. The optimal target mode: (**a**) time−domain waveform and (**b**) SES.

For comparison, the simulated vibration signal $x(t)$ is processed using VMD. In the implementation of VMD, the number of modes is set as 5, and the balancing parameter α value is set as 950, which are the same as those described in the previous analysis of SVMD. The results obtained by VMD are shown in Figure 8. The target mode is also the third mode shown in Figure 9a, and its SES is shown in Figure 9b. On the premise of reasonable parameter settings, VMD can also successfully extract the simulated FCF of 122.667 Hz. As the FCFs extracted by SVMD and VMD have the same value, the PA indices of the corresponding target modes, calculated by Equation (15) or (21), achieve the same value. However, the EC index of the target mode obtained by SVMD is calculated to be 0.7264 according to Equation (14), while the EC index of the target mode obtained by VMD is calculated to be 0.7264. According to the definition of the EC index, the greater value of the EC index means that the fault characteristic component in the target mode is more prominent, and the degree of suppression of interference components with high energy is better. Therefore, in terms of the EC index, the performance of SVMD has certain superiority over VMD. Additionally, as VMD has two key parameters, i.e., the number of modes and balancing parameter α, needing to be set reasonably or optimized, while SVMD has only one, i.e., the balancing parameter α, the implementation of the SVMD method is more efficient than that of the VMD method under the same conditions.

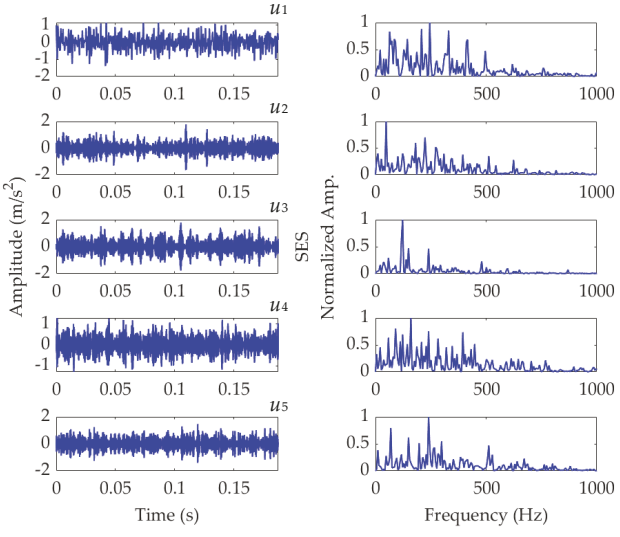


Figure 8. Results of the simulated vibration signal $x(t)$ obtained by VMD.

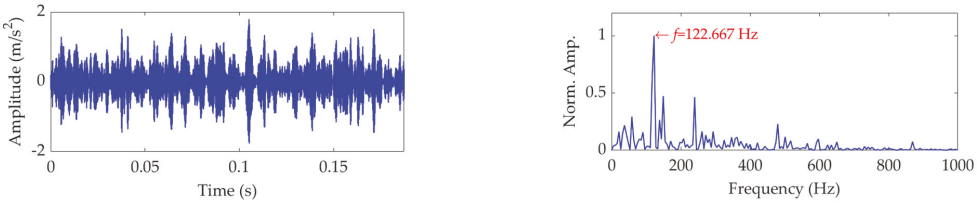


Figure 9. The target mode obtained by VMD: (**a**) time−domain waveform and (**b**) SES.

6. Experimental Evaluation

In this section, the effectiveness of the proposed rolling bearing fault diagnosis method is investigated using three experimental vibration datasets. We focus on the fault feature extraction for the inner race, outer race, and balls of the rolling bearing. Based on this, we use the proposed method to extract the actual FCF from the vibration dataset of the rolling bearing and then compare the possible FCF with the theoretical FCFs of the three components mentioned above. The component whose FCF is closest to the actual FCF can be considered as the faulty component. This is the basic principle of the proposed method for rolling bearing fault diagnosis. According to the previous analysis, we express the theoretical FCFs of the inner race, outer race, and balls of the rolling bearing as f_{ir}, f_{or}, and f_{ba}, respectively. Then, the PA index shown in Equation (15) needs to be specifically modified as Equation (22):

$$PA = \left(|f_{\mathrm{ma}} - f_{\mathrm{ir}}| \cdot |f_{\mathrm{ma}} - f_{\mathrm{or}}| \cdot |f_{\mathrm{ma}} - f_{\mathrm{ba}}|\right)^{1/3} \tag{22}$$

The calculation of the EP index should also be adjusted accordingly.

In the study of this paper, the experimental datasets associated with the rolling bearing come from the Bearing Data Center of Case Western Reserve University [40]. The bearing testbed is shown in Figure 10, which consists of a driving motor (left), a torque transducer/encoder (center), and a dynamometer (right). The driving motor shaft is supported by the test bearing, which was implanted with single point fault using electro-discharge machining. Vibration data associated with the test bearing were collected using accelerometers. Vibration signals were collected using a 16-channel DAT recorder and were post processed in a Matlab environment. The sampling frequency of vibration data was 12,000 Hz. Speed and horsepower data were collected using the encoder and torque transducer, respectively. All the datasets to be analyzed were selected as the vibration datasets of the drive end bearing.

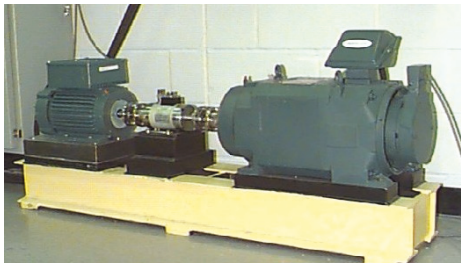


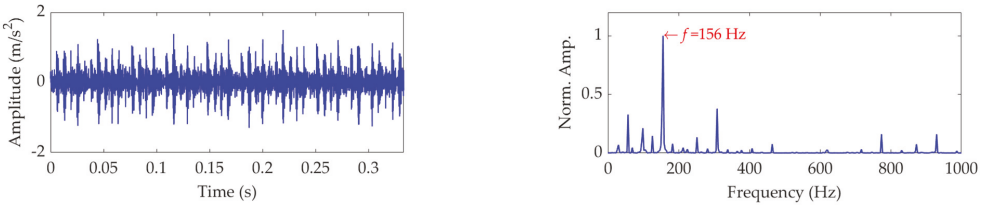
Figure 10. Rolling bearing testbed.

6.1. Inner Race Fault-Related Vibration Dataset Analysis

In this case, we chose the test bearing with a single fault in the inner race, which was 0.007″ in diameter and 0.011″ in depth. The motor load was set as 3 HP. The actual

speed of the motor was 1721 rpm, measured by the encoder, and thus the corresponding rotational frequency was $f_r = 1721/60 = 28.6833$ Hz. Then, the theoretical FCF of the inner race was calculated as $f_{ir} = 5.4152 \cdot f_r = 155.3260$ Hz, the theoretical FCF of the outer race was calculated as $f_{or} = 3.5848 \cdot f_r = 102.8240$ Hz, and the theoretical FCF of the ball was calculated as $f_{ba} = 4.7135 \cdot f_r = 135.1989$ Hz.

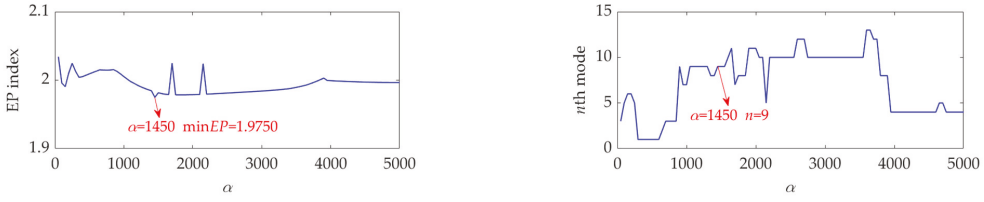
The length of the sampled vibration data associated with the inner race fault was 122,917, and its mean value and standard deviation value were 0.0047 and 0.3136, respectively. A dataset selected from the sampled vibration data is shown in Figure 11a, and its mean value and standard deviation value are 0.0048 and 0.3128 respectively. The SES of the dataset is shown in Figure 11b, in which the actual extracted FCF is 156 Hz, consistent with the theoretical FCF of the inner race. Nevertheless, in order to verify the effectiveness of the proposed method, this vibration dataset was further processed using SVMD.



(**a**) (**b**)

Figure 11. Vibration dataset associated with the inner race fault: (**a**) time−domain waveform and (**b**) SES.

To find an optimal value for the balancing parameter α of SVMD, the α value was set to increase from 50 to 5000 at a step size of 50. In each step of α increase, SVMD was first implemented with the current α value for this vibration dataset to obtain a number of modes, then the EP index of each mode was calculated according to Equation (16), and finally the mode indicated by the minimum EP value was considered as the target mode corresponding to current step. When the α value increased to 5000 and the last target mode was achieved, the relationship curves between the α value and the EP index and order number of the target mode in each step were plotted, as shown in Figure 12. It can be seen that, when the α value is 1450, the EP index achieves an global minimum value, and the corresponding target mode is the ninth mode. Hence, we set the optimal value for the balancing parameter α of SVMD to be $\alpha_{opt} = 1450$ and applied the SVMD algorithm to decompose the vibration dataset in this case. The decomposition results are shown in Figure 13. The variation curve of the EP index for each mode is shown in Figure 14. It can be seen that the ninth mode has the minimum value of the EP index, and it was thus chosen as the optimal target mode, which is consistent with the result in Figure 12b. The optimal target mode containing the complete fault information about the inner race is shown with details in Figure 15a, and its SES is shown in Figure 15b. It can be seen that the actual extracted FCF is 156 Hz, which closely matches the theoretical FCF of the inner race, i.e., 155.3260 Hz, and other frequency components acting as interferences are well suppressed. As noise interference is inevitable in the rolling bearing vibration signal, and the rolling bearing speed is impossible to keep strictly constant due to rolling surface damage, there inevitably exists a difference between the actual extracted FCF and theoretical FCF. However, this difference is very small and perfectly acceptable. The experimental result of this case verifies the effectiveness of the proposed EP index and SVMD method for the rolling bearing inner race fault diagnosis.



(**a**) (**b**)

Figure 12. Relationship between the α value and (**a**) the EP index value and (**b**) order number of the target mode at each step of α increase.

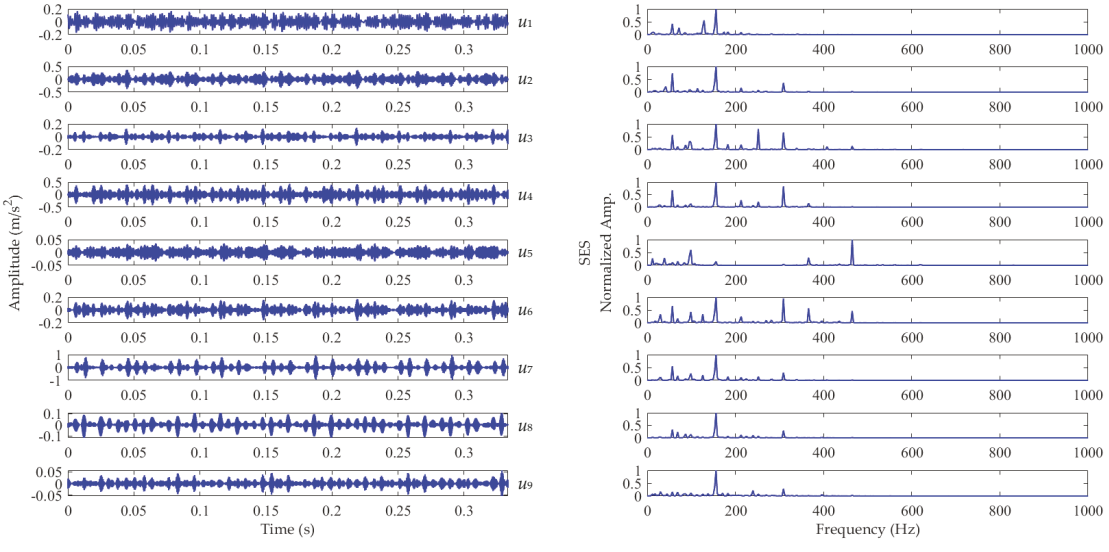


Figure 13. Results of the vibration dataset associated with the inner race fault obtained using SVMD with α_{opt} = 1450.

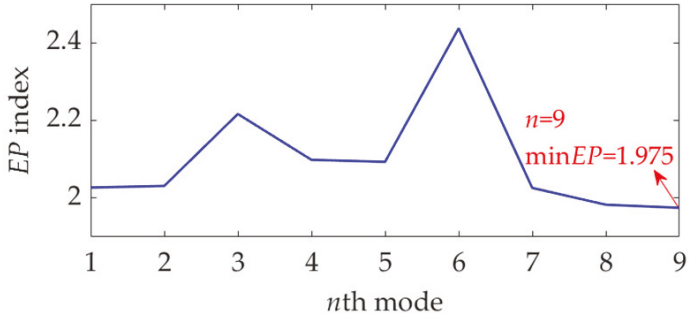
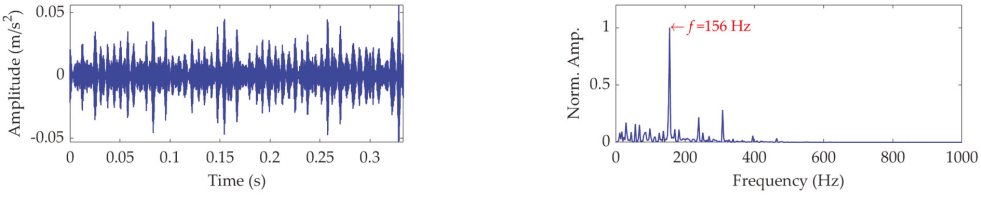


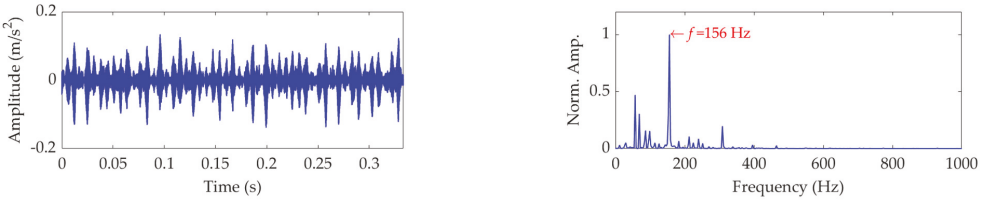
Figure 14. The EP index of each mode.



(**a**) (**b**)

Figure 15. The optimal target mode: (**a**) time−domain waveform and (**b**) SES.

As a comparison, the vibration dataset of this case was decomposed using VMD. In the implementation of VMD, the number of modes was determined as 9 by experiments, and the balancing parameter α was optimally set as 500 using the line search method described previously. Among the modes obtained by the VMD method, the target mode was also the ninth mode, which is specifically shown in Figure 16a, and its SES is shown in Figure 16b. The FCF extracted by VMD is the same as that extracted by SVMD, which means that the PA indices of the target modes extracted by VMD and SVMD achieved the same value according to Equation (22). Nevertheless, the EC index of the target mode extracted by SVMD was calculated to be 0.8147 and was higher than that of the target mode extracted by VMD, which was calculated to be 0.7942. In this scenario, compared with the VMD method, the SVMD method is more effective in high energy interference components' suppression.



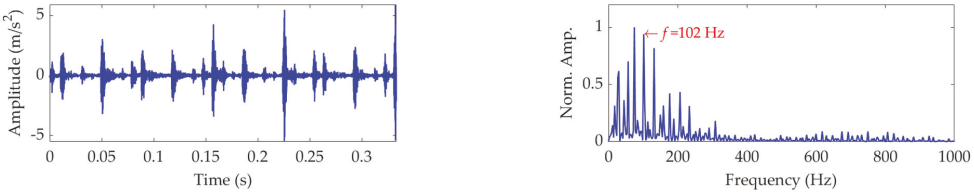
(**a**) (**b**)

Figure 16. The target mode obtained by VMD: (**a**) time−domain waveform and (**b**) SES.

6.2. Outer Race Fault-Related Vibration Dataset Analysis

In this case, the single implanted fault in the outer race of the test bearing was 0.021 in diameter and 0.011 in depth and located at six o'clock. The motor load was set as 3 HP. The measured motor speed was 1721 rpm using the encoder, and thus the corresponding rotational frequency was $f_r = 1721/60 = 28.6833$ Hz. As the test bearing in this case is the same as the previous case, the theoretical FCF of the inner race was $f_{ir} = 155.3260$ Hz, the theoretical FCF of the outer race was $f_{or} = 102.8240$ Hz, and the theoretical FCF of the ball was $f_{ba} = 135.1989$ Hz.

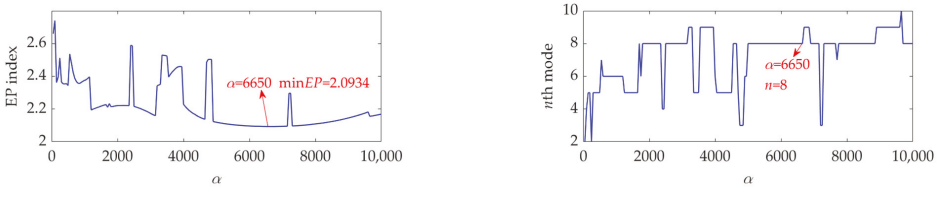
The length of the sampled vibration data with regard to the outer race fault was 121,991 and its mean value and standard deviation value were 0.0035 and 0.5590, respectively. A dataset selected from the sampled vibration data is shown in Figure 17a, and its mean value and standard deviation value are 0.0030 and 0.5511, respectively. The SES of the dataset is shown in Figure 17b. Despite the significant impulsive features in Figure 17a, they do not show obvious periodicity and fail to relate with the FCF of the outer race. In addition, since the fault characteristic component indicated by the extracted actual FCF of 102 Hz is not dominant in the SES, and a lot of strong interference components exist, it is non-rigorous to assert that the outer race is faulty.



(**a**) (**b**)

Figure 17. Vibration dataset associated with the outer race fault: (**a**) time−domain waveform and (**b**) SES.

To apply the SVMD method to the dataset process of this case, the optimal value of the balancing parameter α was searched in the range 50–10,000, and then the α value increased from 50 to 10,000 at a step size of 50. The finally obtained relationship curves between the α value and the EP index and order number of the target mode in each step of α increase are shown in Figure 18. It can be seen that the optimal value of α is α_{opt} = 6650 because the corresponding target mode obtained by SVMD has the global minimum value for the EP index, and this optimal target mode is the eighth mode among the decomposition results. For more details, this dataset was decomposed using SVMD with the obtained optimal α value of 6650. The total number of the decomposed modes is 13, and the EP index of each mode is shown in Figure 19. The eighth mode with the minimum value of the EP index is the target mode, which is also consistent with the result indicated in Figure 18b. The target mode along with its SES is represented in Figure 20. From Figure 20b, we can see that as interference components are greatly removed or suppressed, the actual FCF of 102 Hz is extracted successfully, and the outer race is faulty, since the theoretical FCF of the outer race is 102.8240 Hz. Due to noise interference and very tiny variations of the bearing speed caused by the rolling surface damage, the difference between the actual extracted FCF and theoretical FCF is inevitable but perfectly acceptable. This result suggests that, with the help of the proposed EP index, the target mode can be well selected from the multiple modes obtained by SVMD, and the SVMD method with the optimized balance parameter can effectively extract the fault characteristics of outer race for the rolling bearing.



(**a**) (**b**)

Figure 18. Relationship between the α value and (**a**) the EP index value and (**b**) order number of the target mode at each step of α increase.

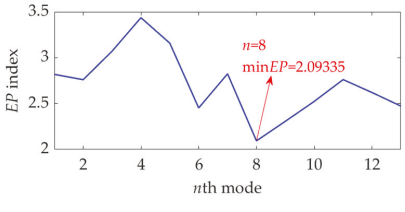
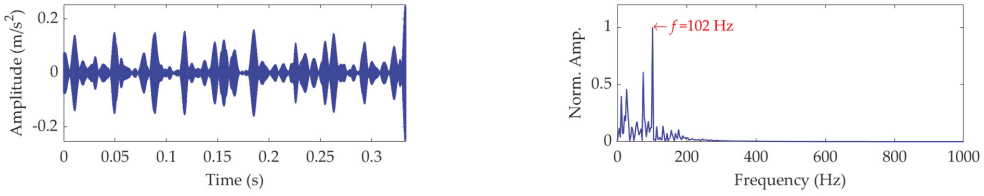


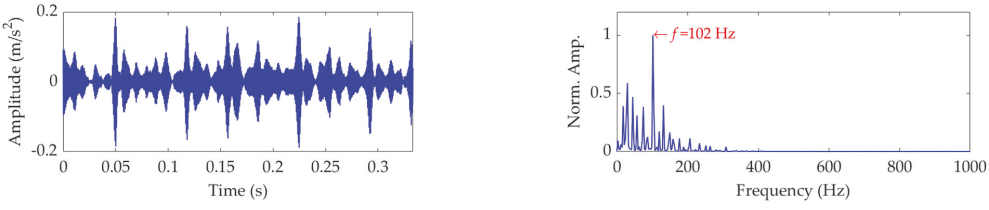
Figure 19. The EP index value of each mode.



(**a**) (**b**)

Figure 20. The optimal target mode: (**a**) time−domain waveform and (**b**) SES.

Now, we further use the VMD method to decompose the dataset in this case. According to the preceding analysis, we set the number of modes as 13 and the balancing parameter as 6650. In fact, this set of values is also a set of relatively optimal values for VMD, which has been validated by the linear search method mentioned previously. Among the results, the target mode is the fourth mode, as Figure 21a shows, and its SES is represented in Figure 21b. Although the VMD method is also able to extract the accurate FCF, in practice, it is cumbersome to acquire a set of relatively optimal values for the number of modes and balancing parameter of VMD. By comparison, the SVMD method is easier to implement. In addition, the EC index values of the target modes extracted by SVMD and VMD were calculated to be 0.6953 and 0.6406, respectively, which indicates that the target mode extracted by SVMD has lower energy interference components and, correspondingly, its fault characteristic component is more prominent.



(**a**) (**b**)

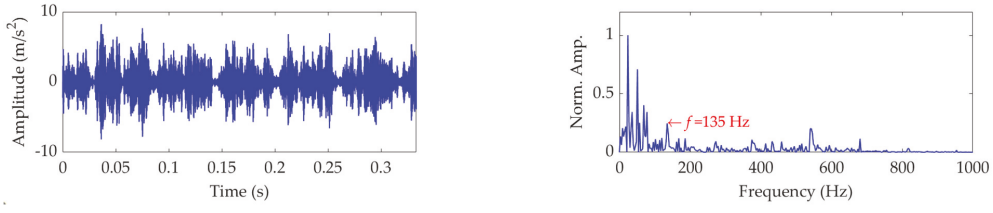
Figure 21. The target mode obtained by VMD: (**a**) time−domain waveform and (**b**) SES.

6.3. Ball Fault-Related Vibration Dataset Analysis

In this case, one of the balls in the test bearing had a single implanted fault with 0.028″ diameter and 0.011″ depth. The motor load was set as 3 HP. The approximate speed of the motor was 1730 rpm, and the corresponding rotational frequency was $f_r = 1730/60 = 28.83$ Hz. Hence, the theoretical FCF of the inner race was calculated as $f_{ir} = 5.4152 \cdot f_r = 156.1202$ Hz, the theoretical FCF of the outer race was calculated as $f_{or} = 3.5848 \cdot f_r = 103.3498$ Hz, and the theoretical FCF of the ball was calculated as $f_{ba} = 4.7135 \cdot f_r = 135.8902$ Hz.

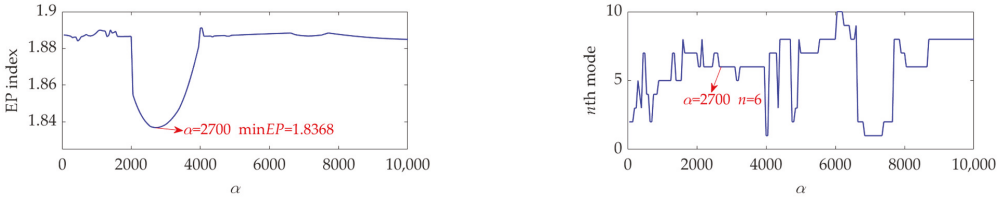
The length of the sampled vibration data corresponding to the ball fault was 120,984, and its mean value and standard deviation value were 0.0190 and 2.1449, respectively. A dataset selected from the sampled vibration data, together with its SES, is shown in Figure 22. The mean value and standard deviation value of the dataset were 0.0142 and 2.1156, respectively. As in the SES, the interference components were dominant and the fault characteristic component itself was quite weak; it is a challenging task to identify the correct FCF and judge that a ball in the rolling bearing is faulty. Hence, we further used the SVMD method to analyze this vibration dataset. To determine a proper value for the balancing parameter α of SVMD, the α value was increased from 50 to 10,000 at a step size of 50, and in each step, the target mode among the decomposed modes by SVMD was selected with the help of the EP index. Eventually, we could obtain the relationship curves between the α value and the EP index, as well as the order number of the target mode in

each step of α increase, all of which are shown in Figure 23. It can be seen that when the α value is 2700, the target mode obtained by SVMD has the global minimum value for the EP index. Therefore, in this case, the optimal value for the balancing parameter α of SVMD is determined as $\alpha_{opt} = 2700$, and the corresponding target mode is the sixth mode indicated in Figure 23b. Then, the vibration dataset of the faulty ball was decomposed by SVMD with the optimal α value of 2700 and the results are shown in Figure 24. The EP index values of all modes are displayed as Figure 25. Hence, the optimal target mode is the sixth mode with the minimum value of the EP index, which is in agreement with the result in Figure 23b. To be specific, the optimal target mode and its SES are shown in Figure 26. It can be seen that the fault characteristic component is the strongest, and the FCF of the ball, i.e., 135 Hz, can be easily identified, which is very close to the theoretical value of 135.8902 Hz. Considering the effect of random slippage of the balls, as well as the noise interference and the slight fluctuation of the bearing speed caused by rolling surface damage, there inevitably exists a difference between the actual FCF and theoretical FCF of the ball. Nevertheless, the actual extracted FCF is accurate enough and acceptable. Such a result shows that, guided by the proposed EP index, the SVMD method can effectively extract the useful fault characteristic information from rolling bearing vibration signal with strong interferences.



(**a**) (**b**)

Figure 22. Vibration dataset associated with the ball fault: (**a**) time−domain waveform and (**b**) SES.



(**a**) (**b**)

Figure 23. Relationship between the α value and (**a**) the EP index value and (**b**) order number of the target mode at each step of α increase.

Lastly, the vibration dataset of this case was also decomposed by the VMD method for comparison. The balancing parameter α of VMD was set as the optimal value of 2700 used in the SVMD method. The number of modes was set as 20, which is an optimal value determined by multiple experiments. In the results obtained by VMD, the target mode was the 18th mode, whose waveform and SES are shown in Figure 27. It can be seen that, as long as the parameters are set properly, VMD is also able to extract the FCF of the rolling bearing ball from the vibration dataset. However, it is important to note that, in a practical application, the two key parameters of VMD are not easy to determine. As only one key parameter, i.e., the balancing parameter α, needs to be optimized in SVMD, the SVMD method shows higher efficiency than the VMD method under the same conditions. For further comparison, the EC indices of the target modes extracted by SVMD and VMD were calculated according to Equation (14) and achieved the values of 0.5702 and 0.4793,

respectively. In this case, the target mode extracted by SVMD also has a better EC index and more prominent fault characteristic component.

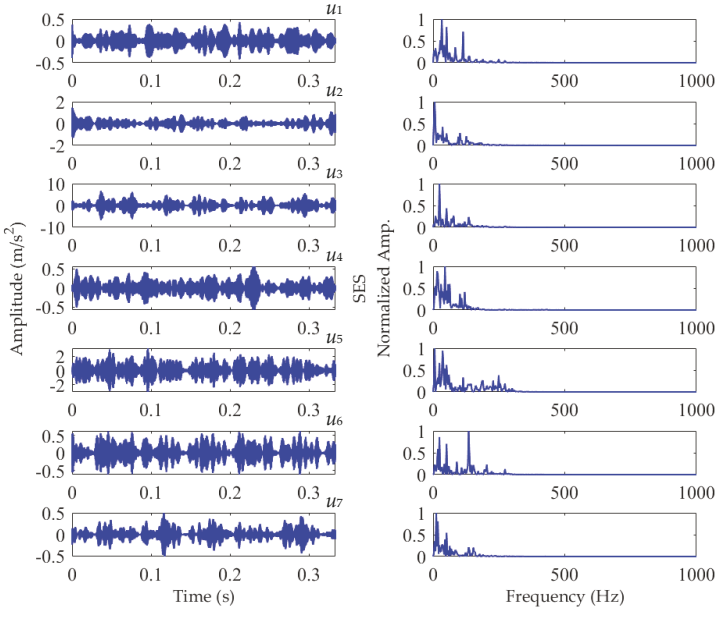


Figure 24. Results of the vibration dataset associated with the ball fault obtained using SVMD with α_{opt} = 2700.

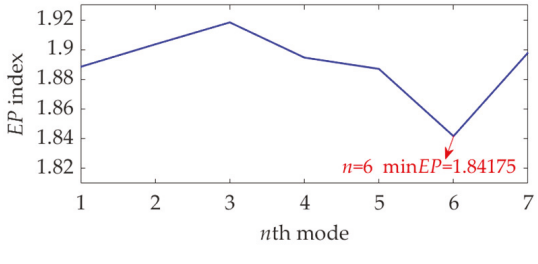
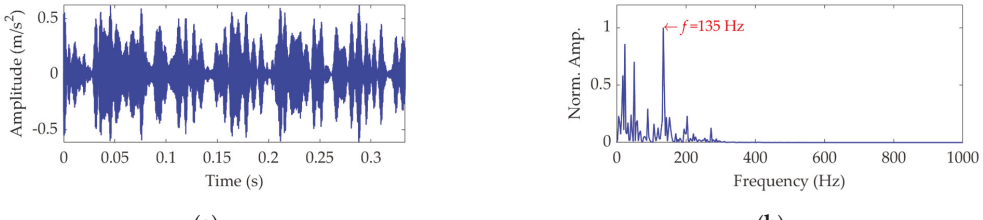
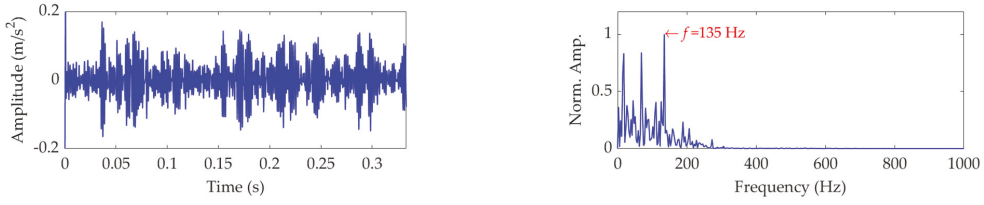


Figure 25. The EP index value of each mode.



(**a**) (**b**)

Figure 26. The optimal target mode: (**a**) time−domain waveform and (**b**) SES.



(**a**) (**b**)

Figure 27. The target mode obtained by VMD: (**a**) time−domain waveform and (**b**) SES.

7. Conclusions

In this paper, we propose a rolling bearing fault diagnosis method based on SVMD and the EP index. As there is only the balancing parameter needing to be optimized for SVMD, the SVMD method is more feasible to implement than the VMD method, which needs to optimize the number of modes and balancing parameter simultaneously. Nevertheless, it is not easy to determine the optimal value of the balancing parameter for SVMD, and the target mode containing ideal fault characteristic information is difficult to select from the multiple modes obtained by SVMD. In view of the existing shortcomings of SVMD, which are also true in VMD, the new proposed EP index can effectively indicate the target mode from the results of SVMD. Accordingly, an optimal value for the balancing parameter of SVMD can be easily achieved using the line search method guided by the EP index. The simulation and experimental results verify the effectiveness and practicability of the EP index and also demonstrate that the SVMD method has strong anti-noise and anti-interference ability, and thus can successfully extract the fault feature from vibration signal to realize the rolling bearing fault diagnosis. In addition, quantified by the new proposed EC index, the SVMD method shows better performance in interference suppression and fault feature enhancement than the VMD method. In the future, we may apply the SVMD method and EP index to fault feature extraction and fault diagnosis for a multistage gearbox, especially the wind turbine gearbox.

Author Contributions: Writing—original draft, methodology, Y.G.; supervision, Y.Y.; formal analysis, S.J.; validation, X.J.; resources, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 51975517), the Key R & D Program of Zhejiang Province (Grant No. 2021C01086), the Ningbo Natural Science Foundation of China (Grant No. 2021J038) and the Innovative Experiment Project of Zhejiang University of Technology (No. PX-79192739).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Beretta, M.; Julian, A.; Sepulveda, J.; Cusidó, J.; Porro, O. An ensemble learning solution for predictive maintenance of wind turbines main bearing. *Sensors* **2021**, *21*, 1512. [CrossRef]
2. Huang, Y.; Lin, J.; Liu, Z.; Wu, W. A modified scale-space guiding variational mode decomposition for high-speed railway bearing fault diagnosis. *J. Sound Vib.* **2019**, *444*, 216–234. [CrossRef]
3. Elasha, F.; Li, X.; Mba, D.; Ogundare, A.; Ojolo, S. A novel condition indicator for bearing fault detection within helicopter transmission. *J. Vib. Eng. Technol.* **2021**, *9*, 215–224. [CrossRef]
4. Wang, X.; Lu, S.; Chen, K.; Wang, Q.; Zhang, S. Bearing fault diagnosis of switched reluctance motor in electric vehicle powertrain via multisensor data fusion. *IEEE Trans. Ind. Inform.* **2021**, *18*, 2452–2464. [CrossRef]

5. Cerrada, M.; Sánchez, R.V.; Li, C.; Pacheco, F.; Cabrera, D.; de Oliveira, J.V.; Vásquez, R.E. A review on data-driven fault severity assessment in rolling bearings. *Mech. Syst. Signal Process.* **2018**, *99*, 169–196. [CrossRef]
6. Pathiran, A.R.; Erikiananda, K.; Getachew, T.; Gziabher, H.G. Performance and predict the ball bearing faults using wavelet packet decomposition and ANFIS. *Int. J. Eng. Sci.* **2019**, *11*, 33–47. [CrossRef]
7. Wang, J.; Du, G.; Zhu, Z.; Shen, C.; He, Q. Fault diagnosis of rotating machines based on the EMD manifold. *Mech. Syst. Signal Process.* **2020**, *135*, 106443. [CrossRef]
8. Ye, X.; Hu, Y.; Shen, J.; Chen, C.; Zhai, G. An adaptive optimized TVF-EMD based on a sparsity-impact measure index for bearing incipient fault diagnosis. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–11. [CrossRef]
9. Yasir, M.N.; Koh, B.H. Data decomposition techniques with multi–scale permutation entropy calculations for bearing fault diagnosis. *Sensors* **2018**, *18*, 1278. [CrossRef]
10. Li, X.; Ma, J.; Wang, X.; Wu, J.; Li, Z. An improved local mean decomposition method based on improved composite interpolation envelope and its application in bearing fault feature extraction. *ISA Trans.* **2020**, *97*, 365–383. [CrossRef]
11. Xu, Y.; Tian, W.; Zhang, K.; Ma, C. Application of an enhanced fast kurtogram based on empirical wavelet transform for bearing fault diagnosis. *Meas. Sci. Technol.* **2019**, *30*, 035001. [CrossRef]
12. Xu, Y.; Deng, Y.; Zhao, J.; Tian, W.; Ma, C. A novel rolling bearing fault diagnosis method based on empirical wavelet transform and spectral trend. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 2891–2904. [CrossRef]
13. Zhao, H.; Zuo, S.; Hou, M.; Liu, W.; Yu, L.; Yang, X.; Deng, W. A novel adaptive signal processing method based on enhanced empirical wavelet transform technology. *Sensors* **2018**, *18*, 3323. [CrossRef]
14. Liu, Q.; Yang, J.; Zhang, K. An improved empirical wavelet transform and sensitive components selecting method for bearing fault. *Measurement* **2022**, *187*, 110348. [CrossRef]
15. Li, H.; Liu, T.; Wu, X.; Chen, Q. An optimized VMD method and its applications in bearing fault diagnosis. *Measurement* **2020**, *166*, 108185. [CrossRef]
16. Ding, J.; Huang, L.; Xiao, D.; Li, X. GMPSO-VMD algorithm and its application to rolling bearing fault feature extraction. *Sensors* **2020**, *20*, 1946. [CrossRef]
17. Zhang, M.; Jiang, Z.; Feng, K. Research on variational mode decomposition in rolling bearings fault diagnosis of the multistage centrifugal pump. *Mech. Syst. Signal Process.* **2017**, *93*, 460–493. [CrossRef]
18. Dibaj, A.; Hassannejad, R.; Ettefagh, M.M.; Ehghaghi, M.B. Incipient fault diagnosis of bearings based on parameter–optimized VMD and envelope spectrum weighted kurtosis index with a new sensitivity assessment threshold. *ISA Trans.* **2021**, *114*, 413–433. [CrossRef]
19. Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [CrossRef]
20. Jiang, X.; Wang, J.; Shi, J.; Shen, C.; Huang, W.; Zhu, Z. A coarse-to-fine decomposing strategy of VMD for extraction of weak repetitive transients in fault diagnosis of rotating machines. *Mech. Syst. Signal Process.* **2019**, *116*, 668–692. [CrossRef]
21. Zhou, X.; Li, Y.; Jiang, L.; Zhou, L. Fault feature extraction for rolling bearings based on parameter–adaptive variational mode decomposition and multi-point optimal minimum entropy deconvolution. *Measurement* **2021**, *173*, 108469. [CrossRef]
22. Ying, W.; Zheng, J.; Pan, H.; Liu, Q. Permutation entropy–based improved uniform phase empirical mode decomposition for mechanical fault diagnosis. *Digit. Signal Process.* **2021**, *117*, 103167. [CrossRef]
23. Zheng, J.; Su, M.; Ying, W.; Tong, J.; Pan, Z. Improved uniform phase empirical mode decomposition and its application in machinery fault diagnosis. *Measurement* **2021**, *179*, 109425. [CrossRef]
24. Liang, T.; Lu, H. A novel method based on multi-island genetic algorithm improved variational mode decomposition and multi–features for fault diagnosis of rolling bearing. *Entropy* **2020**, *22*, 995. [CrossRef] [PubMed]
25. Zhang, C.; Wang, Y.; Deng, W. Fault diagnosis for rolling bearings using optimized variational mode decomposition and resonance demodulation. *Entropy* **2020**, *22*, 739. [CrossRef]
26. Wang, Q.; Yang, C.; Wan, H.; Deng, D.; Nandi, A.K. Bearing fault diagnosis based on optimized variational mode decomposition and 1D convolutional neural networks. *Meas. Sci. Technol.* **2021**, *32*, 104007. [CrossRef]
27. Shi, W.; Wen, G.; Huang, X.; Zhang, Z.; Zhou, Q. The VMD-scale space based hoyergram and its application in rolling bearing fault diagnosis. *Meas. Sci. Technol.* **2020**, *31*, 125006. [CrossRef]
28. Wang, J.; Zhan, C.; Li, S.; Zhao, Q.; Liu, J.; Xie, Z. Adaptive variational mode decomposition based on Archimedes optimization algorithm and its application to bearing fault diagnosis. *Measurement* **2022**, *191*, 110798. [CrossRef]
29. Gai, J.; Shen, J.; Hu, Y.; Wang, H. An integrated method based on hybrid grey wolf optimizer improved variational mode decomposition and deep neural network for fault diagnosis of rolling bearing. *Measurement* **2020**, *162*, 107901. [CrossRef]
30. Jin, Z.; He, D.; Wei, Z. Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN. *Eng. Appl. Artif. Intel.* **2022**, *110*, 104713. [CrossRef]
31. Yan, X.; Xu, Y.; She, D.; Zhang, W. A bearing fault diagnosis method based on PAVME and MEDE. *Entropy* **2021**, *23*, 1402. [CrossRef] [PubMed]
32. Yan, X.; Jia, M. Application of CSA-VMD and optimal scale morphological slice bispectrum in enhancing outer race fault detection of rolling element bearings. *Mech. Syst. Signal Process.* **2019**, *122*, 56–86. [CrossRef]
33. He, D.; Liu, C.; Jin, Z.; Ma, R.; Chen, Y.; Shan, S. Fault diagnosis of flywheel bearing based on parameter optimization variational mode decomposition energy entropy and deep learning. *Energy* **2022**, *239*, 122108. [CrossRef]

34. Nazari, M.; Sakhaei, S.M. Variational mode extraction: A new efficient method to derive respiratory signals from ECG. *IEEE J. Biomed. Health* **2017**, *22*, 1059–1067. [CrossRef] [PubMed]
35. Guo, Y.; Jiang, S.; Yang, Y.; Jin, X.; Wei, Y. Gearbox fault diagnosis based on improved variational mode extraction. *Sensors* **2022**, *22*, 1779. [CrossRef]
36. Nazari, M.; Sakhaei, S.M. Successive variational mode decomposition. *Signal Process.* **2020**, *174*, 107610. [CrossRef]
37. Borghesani, P.; Pennacchi, P.; Chatterton, S. The relationship between kurtosis-and envelope-based indexes for the diagnostic of rolling element bearings. *Mech. Syst. Signal Process.* **2014**, *43*, 25–43. [CrossRef]
38. Liang, M.; Bozchalooi, I.S. An energy operator approach to joint application of amplitude and frequency- demodulations for bearing fault detection. *Mech. Syst. Signal Process.* **2010**, *24*, 1473–1494. [CrossRef]
39. Buzzoni, M.; D'Elia, G.; Cocconcelli, M. A tool for validating and benchmarking signal processing techniques applied to machine diagnosis. *Mech. Syst. Signal Process.* **2020**, *139*, 106618. [CrossRef]
40. Case Western Reserve University Bearing Data Center. Available online: https://engineering.case.edu/bearingdatacenter (accessed on 26 February 2022).

sensors

MDPI

Article

Compression Reconstruction and Fault Diagnosis of Diesel Engine Vibration Signal Based on Optimizing Block Sparse Bayesian Learning

Huajun Bai [1], Liang Wen [1,2], Yunfei Ma [1] and Xisheng Jia [1,*]

[1] Shijiazhuang Campus, Army Engineering University of PLA, Shijiazhuang 050003, China; bai_huajun@sina.com (H.B.); lwenmark@163.com (L.W.); fcz1992@sina.com (Y.M.)
[2] Hebei Key Laboratory of Condition Monitoring and Assessment of Mechanical Equipment, Shijiazhuang 050003, China
* Correspondence: xs_jia@sina.cn

Abstract: It is critical to deploy wireless data transmission technologies remotely, in real-time, to monitor the health state of diesel engines dynamically. The usual approach to data compression is to collect data first, then compress it; however, we cannot ensure the correctness and efficiency of the data. Based on sparse Bayesian optimization block learning, this research provides a method for compression reconstruction and fault diagnostics of diesel engine vibration data. This method's essential contribution is combining compressive sensing technology with fault diagnosis. To achieve a better diagnosis effect, we can effectively improve the wireless transmission efficiency of the vibration signal. First, the dictionary is dynamically updated by learning the dictionary using singular value decomposition to produce the ideal sparse form. Second, a block sparse Bayesian learning boundary optimization approach is utilized to recover structured non-sparse signals rapidly. A detailed assessment index of the data compression effect is created. Finally, the experimental findings reveal that the approach provided in this study outperforms standard compression methods in terms of compression efficiency and accuracy and its ability to produce the desired fault diagnostic effect, proving the usefulness of the proposed method.

Keywords: diesel engine; data compression; vibration signal; K-SVD; fault diagnosis

Citation: Bai, H.; Wen, L.; Ma, Y.; Jia, X. Compression Reconstruction and Fault Diagnosis of Diesel Engine Vibration Signal Based on Optimizing Block Sparse Bayesian Learning. *Sensors* **2022**, *22*, 3884. https://doi.org/10.3390/s22103884

Academic Editor: Yolanda Vidal

Received: 12 April 2022
Accepted: 17 May 2022
Published: 20 May 2022

1. Introduction

Diesel engines have been widely used in energy, construction machinery, and military equipment. Vibration signals are transmitted dynamically and synchronously in real-time, Playing a pivotal role in real-time online monitoring of diesel engine health [1–3]. It can effectively reduce the incidence of equipment failure, downtime, and management costs. Compared with traditional wired data transmission, edge computing wireless data transmission methods can significantly improve the real-time, flexibility, and ease of use of data, according to the Nyquist sampling theorem [4]. To realize the collection of high-frequency vibration signals of the equipment, we will inevitably generate a large amount of data. However, big data is constrained by network bandwidth during wireless transmission. Whether it can support the problems of massive data, high concurrency, low latency, and low power consumption is yet to be determined.

Recently, it has become a research hotspot that researchers focus on. For example, Antonopoulos et al. [5] embedded compression algorithms into hardware systems to improve the work efficiency of transmitting large amounts of data wirelessly. Ma et al. [6] used a distributed video codec scheme to enhance the processing power of a single node for traditional data compression. Yi et al. [7] proposed an adaptive data compression and transmission range extension scheme to improve the data collection rate of sink nodes. Hameed et al. [8] used lossless compression technology and Huffman coding encryption

technology to provide effective means for remote monitoring security and compressibility of electrocardiography (ECG) data. Therefore, before the data are wirelessly transmitted, real-time synchronous sampling and compression of the original vibration data is the best solution to solve the above problems.

Compressive sensing (CS) is a new technical theory that has emerged in recent years [9]. Due to its outstanding performance in data compression and reconstruction, it has been widely used in the field of image and sound. Use the observation matrix to map the original vibration signal from the high-dimensional space to the low-dimensional space. Then, the original signal is recovered with a high probability from fewer observations through an optimization algorithm. Currently, commonly used compression and re-construction algorithms include greedy algorithm [10], convex optimization algorithm [11], Bayesian learning [12], etc. For example, Liu et al. [13] used a low-pass filtering method to optimize the electrographic signal and used basis pursuit (BP) algorithm to compress and reconstruct the electrocardiogram signal. Cheng et al. [14] used an improved orthogonal matching pursuit (OMP) algorithm to improve seismic data's reconstruction speed and compression effect. Sajjad et al. [15] used a genetic algorithm to optimize the sparse signal and the regularized orthogonal matching pursuit (ROMP) algorithm to reconstruct the image signal. Generally, reciprocating mechanical vibration signals have sparse, non-sparse, and unique structural features. The traditional compression and reconstruction algorithm is used to recover sparse signals with high accuracy and versatility in the above research. However, this type of algorithm only considers its sparsity and is not necessarily suitable for reconstructing reciprocating mechanical vibration signals. Improving the recovery accuracy of structured non-sparse signals becomes crucial.

In the existing Bayesian algorithm, the block sparse Bayesian learning bound optimization (BSBL-BO) algorithm [16] has the potential to solve the problem of structured non-sparse signal reconstruction. The algorithm effectively uses the intra-block correlation of vibration signals to restore structured non-sparse signals. Compared with other traditional compression and reconstruction algorithms, the BSBL-BO algorithm has the advantages of high signal recovery accuracy and good compression effect and has been widely used in electrocardiograms and radar. For example, Mahrous et al. [17] proposed a space-time sparse Bayesian learning method. By optimizing the BSBL-BO algorithm, the compression and reconstruction of multi-channel electro-encephalogram (EEG) signals are realized. Li et al. [18] used an enhanced narrow-band interference separation algorithm for radar to achieve compression and reconstruction of radar signals through the BSBL framework, proving the feasibility of the BSBL-BO algorithm for data compression. However, this algorithm has not been studied much in reciprocating mechanical vibration signals in previous studies. This paper carries out related research based on the BSBL-BO algorithm to fill the gap.

An essential prerequisite for CS is the sparsity in the original vibration signal. Sparsity plays a crucial role in the accuracy of the reconstruction of recovered data. Therefore, an efficient data dictionary is needed to improve the signal's sparsity. Classical dictionaries include discrete cosine transform (DCT) [19], discrete Fourier transform (DFT) [20], and wavelet packet transform (discrete wavelet transform, DWT) [21] are fixed dictionaries. The ideal sparse representation can only be obtained when the atomic features in this dictionary type are the same as the original vibration information. There is also a dictionary, commonly used K-singular value decomposition (K-SVD) [22] and optimal directions (method of optimal directions, MOD) [23]. The dictionary is dynamically updated through training to obtain the optimal sparse representation. Compared with the fixed dictionary, it has the advantage of solid adaptive ability. For example, Li et al. [24] used the K-SVD algorithm to update the dictionary to improve the sparsity of image signals. Yang et al. [25] used the K-SVD algorithm to enhance the sparse representation of medical images to obtain better compression and reconstruction accuracy.

Diesel engines often have various failures in their daily work. Among them, the loss of the diesel engine refers to the phenomenon of increased valve clearance, severe deformation

of a valve seat ring, burning oil, and severe wear of piston rings during operation. As a result, the diesel engine cannot work normally, and there is a more significant safety hazard. To reduce the occurrence rate of diesel engine failures and improve stability and safety, researchers have carried out a great deal of research work and achieved fruitful research results. Gu et al. [26] applied the multivariate empirical mode decomposition to the fault diagnosis of diesel engine misfire and achieved good fault classification results by using the SVM classifier. Chen et al.'s [27] harmony search optimizer is used to set hyper-parameters of the variational stacked autoencoder. This method has been well applied in the fault detection of diesel engines. Wang et al. [28] proposed the plan of particle swarm optimization probabilistic neural network (probabilistic neural network, PNN) and support vector machine. Effective diagnosis of common engine failures is achieved. In recent years, the application of compressed sensing theory to fault diagnosis has gradually attracted the attention of researchers, and some research results have been completed. Zhang et al. [29] trained several over-complete dictionaries with a dictionary learning method. Thereby, redundant dictionaries corresponding to different fault categories are obtained. The matching tracking algorithm is used to determine. The error of the reconstructed signal under various dictionaries is compared to realize the diagnosis of the fault category. Tang et al. [30] first obtained the compressed acquisition signal. Then, given the specified sparsity, the matching pursuit algorithm is used to directly obtain the first few fault characteristic frequencies with enormous energy. To realize the identification and diagnosis of fault signals, Du et al. [31] used a dictionary constructed from Fourier transform matrices. The fault features are directly extracted in the compressed measurement domain to realize fault diagnosis of vibration signals.

Although compression technology has been widely used, there are still the following problems or deficiencies:

1. In the process of wireless transmission, due to the limitation of network bandwidth and low power consumption, massive vibration signals bring considerable challenges to data storage and wireless network transmission;
2. The problem of the reconstruction accuracy of the structured non-sparse signal of the reciprocating mechanical vibration signal cannot be satisfied by the traditional data compression technology;
3. Aiming at the compression and reconstruction effects of reciprocating mechanical vibration signals, there is a lack of an effective, comprehensive evaluation index for data compression effects;
4. There is a lack of relevant research on compressive sensing technology and fault diagnosis methods and their application in fault diagnosis of reciprocating machinery.

Using the BSBL-BO algorithm can effectively solve the problem of structured non-sparse signal reconstruction. At the same time, the sparsity of the signal can also be enhanced by the adaptive dynamic updating of the K-SVD dictionary. Combining the two methods can efficiently and accurately recover structured non-sparse signals. Therefore, this paper proposes a compression and reconstruction method based on the BSBL-BO algorithm and the K-SVD dictionary. In addition, this article also establishes an evaluation index for the effect of data compression. First, divide the original signal into blocks. Use the K-SVD dictionary to obtain optimal sparse decomposition to train the actual movement to improve the re-construction performance of the restored signal. Second, use the BSBL-BO algorithm to restore structured non-sparse signals. Compared with other reconstruction algorithms, it has the advantages of high accuracy and a good data compression effect. Finally, the proposed BSBL-KSVD algorithm is verified through a diesel engine valve clearance experiment and fault classification. The experimental results prove that the BSBL-KSVD algorithm proposed in this paper is practical and feasible, providing a reference basis for wireless data transmission of reciprocating mechanical vibration signals.

The main contributions of this paper are summarized as follows:

1. Using the BSBL-KSVD algorithm and exploiting the intra-block correlation of the vibration signal, we can recover the structured non-sparse signal efficiently. Compared

with other traditional compression and reconstruction algorithms. We can effectively improve the reconstruction accuracy and compression effect;
2. A comprehensive evaluation index of compression effect suitable for reciprocating mechanical vibration signal is constructed, and it has a good engineering application prospect;
3. We apply compressed sensing technology to fault diagnosis. The wireless transmission efficiency of the vibration signal can be effectively improved to achieve a better diagnosis effect and has a better reference value.

The second section of this article describes the diesel engine compression reconstruction method model based on BSBL-KSVD; the third part is the comprehensive evaluation index of vibration data compression effect; the fourth part verifies the effectiveness of the compression reconstruction method through preset failure experiments. Finally, this research is summarized.

2. Model of Diesel Engine Compression Reconstruction Method Based on BSBL-KSVD

2.1. Compressed Sensing

In traditional data acquisition and transmission, the Nyquist sampling theorem is used. Usually, the sampling frequency is set to more than twice the highest frequency in the signal under test. Due to the high sampling frequency, a large amount of data is generated. This brings considerable challenges to the wireless data transmission, storage, and remote real-time dynamic monitoring of the operational status of the diesel engine. The emergence of CS theory breaks through the limitation of the traditional vibration signal sampling theorem. Combining the acquisition of vibration signals with the compression process, a small number of signals contains most of the valuable data. Assuming the original signal $x \in R^{N\times 1}$ and observation matrix $\Phi \in R^{M\times N}$ ($M \ll N$), then the signal x is linearly projected on the matrix $y \in R^{M\times 1}$ as a compressed signal. Then, the compressed observation of the original signal $x \in R^{N\times 1}$ can be obtained [32]:

$$y = \Phi x + v \tag{1}$$

Among them, v represents the unknown noise vector. The CS algorithm uses the compressed data y and the measurement matrix Φ to restore the original vibration signal x.

2.2. Block Sparse Bayesian Learning Reconstruction Algorithm

We were using the block structure characteristics of sparse signals. Based on the block sparse Bayesian learning framework, data compression can be realized. In actual engineering applications, the signal x has a block structure feature, as shown in the following equation [16]:

$$X = [\underbrace{x_1, \cdots, x_{d_1}}_{X_1^T}, \cdots, \underbrace{x_{d_{g-1}+1}, \cdots, x_{d_g}}_{X_g^T}]^T \tag{2}$$

The model combined by Equations (1) and (2) is called a block sparse data compression model. We use the characteristics of intra-block correlation to improve the ability of compressed data recovery. Therefore, based on the model in the BSBL framework, it is assumed that the independent x_i between each block satisfies a multivariate Gaussian distribution [16]:

$$p(x_i; \gamma_i, B_i) \sim N(0, \gamma_i B_i), i = 1, \cdots, g \tag{3}$$

Among them, γ_i and B_i both represent unknown parameter variables. γ_i represents a non-negative parameter variable that controls the block sparsity of the original signal x. B_i represents a positive definite matrix used to obtain the related structure between elements in each block. Assuming that the noise vector obeys the Gaussian prior distribution

$p(v;\lambda) \sim N(0,\lambda I)$, use Bayesian principle to obtain the posterior probability of x, as shown in the following equation [16]:

$$p(x\Big|y;\lambda,\{\gamma_i,B_i\}_{i=1}^{g}) \sim N(\mu_x,\sum x) \tag{4}$$

Among them, $\mu_x = \sum_0 \Phi^T(\lambda I + \Phi\sum_0 \Phi^T)^{-1} y$, $\sum_x = (\sum_0^{-1} + \frac{1}{\lambda}\Phi^T\Phi)^{-1}$.

When the parameters λ and $\{\gamma_i, B_i\}_{i=1}^{g}$ are solved, then the maximum posterior estimate of x can be obtained as $\hat{x}$. Next, use the second type of maximum likelihood estimation method to obtain this parameter, as shown in the following equation [16]:

$$\begin{aligned} L(\Theta) &\triangleq -2\log\int p(y|x;\lambda)\,p(x;\{\gamma_i,B_i\}i)dx \\ &= \log|\lambda I + \Phi\sum_0 \Phi^T\Big| + y^T(\lambda I + \Phi\sum_0 \Phi^T)^{-1} y \end{aligned} \tag{5}$$

where Θ represents the parameters λ, $\{\gamma_i, B_i\}_{i=1}^{g}$.

2.3. K-SVD Adaptive Over-Complete Dictionary

The traditional fixed dictionary has a particular sparse representation when the signal is sparsely decomposed. Since the sparse representation of the limited dictionary is unknown, its suitability and flexibility are not strong enough. To further improve the sparsity, we need to use an adaptive dictionary learning method for optimization. Therefore, the K-SVD learning dictionary is used as the spare base to obtain a better sparse representation. The dictionary atom is dynamically updated through training until an adaptive over-complete dictionary is obtained. To ensure that the atomic scale in the dictionary is closer to the atomic scale in the original signal, the training process of dictionary D is expressed as [33]:

$$\min_D \| Y - DA \|_F^2 \text{ s.t. } \| a_i \|_0 \le T \tag{6}$$

In the above equation, Y represents the given training dictionary matrix, A represents a sparse matrix, and T represents the sparsity of the sparse representation vector to be solved.

Initialization D belongs to a super-complete dictionary, and there is a certain degree of redundancy. Suppose that when we update the j-th column atom in dictionary D, we also let E_i be the calculation error after removing the i-th atom; d_j represents the j-th column of dictionary D, and a^i represents the i-th row of sparse matrix A. Then, the objective function is as follows [33]:

$$\| Y - DA \|_F^2 = \left\| Y - \sum_{j=1}^{K} d_j a^i \right\|_F^2 = \left\| (Y - \sum_{j\neq 1} d_j a^j) - d_i a^i \right\|_F^2 = \left\| E_i - d_i a^i \right\|_F^2 \tag{7}$$

When directly decomposing E_i, the elements in the obtained a^i may not be sparse. Therefore, only the non-zero elements in a^i need to be updated, defined as the following equation:

$$w_i = \left\{ k \Big| a^i(k) \neq 0 \right\} \tag{8}$$

represents the index collection of the index of the non-zero element in ai. The SVD decomposition method is used to update the atomic vector gradually, and the sparse representation coefficient matrix A in the dictionary D. Next, we generate a new dictionary through multiple iterative updates.

2.4. Basic Flow of BSBL-KSVD Algorithm

The algorithm flow of compression and reconstruction of diesel engine vibration signal based on BSBL-KSVD is shown in Figure 1. The algorithm mainly includes dictionary training, data compression, and signal reconstruction.

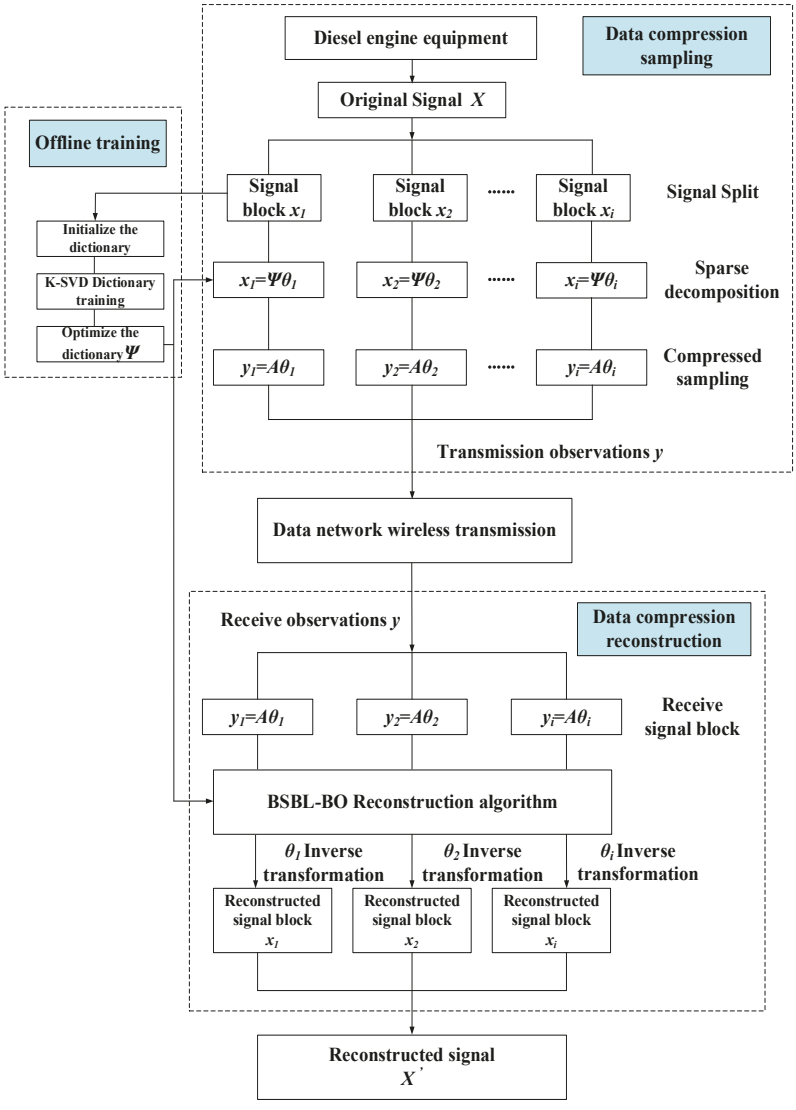


Figure 1. BSBL-KSVD compression reconstruction algorithm flow.

The specific implementation steps are as follows:

Step 1. The signal is divided into blocks. Customize the collected original vibration signal x into i blocks and the size of the elements in each block;

Step 2. Dictionary training: Initialize the dictionary parameters, set the number of training samples, use the K-SVD algorithm to train the examples, and obtain an optimized dictionary Ψ;

Step 3. Data compression: The vibration signal of reciprocating machinery is more complicated than that of rotating machinery. To further improve the sparsity of the signal, the optimized dictionary Ψ can map the signal to the sparse transformation, and the original signal $x = \Psi\theta$ can obtain the sparse transformation signal θ. The sensing matrix $A = \Phi\Psi$

(that is, observation matrix × sparse matrix) compresses the sparse signal data and obtains the data compressed signal observation value $y = A\theta$;

Step 4. Signal transmission: The block-compressed signals $y_1, \ldots, y_i$ are successively transmitted through the data network;

Step 5. Signal reconstruction: After receiving the compressed signal block, using the BSBL-KSVD reconstruction algorithm proposed in this article through the sensor matrix $A_1, \ldots, A_i$ and compressed signal $y_1, \ldots, y_i$ to reconstruct, we obtain the restored sparse signal $\theta_1, \ldots, \theta_i$. At the same time, we perform inverse sparse transformation to obtain reconstructed signal blocks $x_1, \ldots, x_i$ and connect the reconstructed signal blocks one by one and finally form a complete reconstructed signal x'.

The pseudo code of the algorithm (Algorithm 1) is as follows:

Algorithm 1 BSBL-KSVD algorithm pseudo code

1. Input: $x = [x_1, x_2, \ldots, x_i]$, blkLen, N, M;
2. Initialize dictionary parameters: *param. L* = 5, *param. K* = 70, *param. numIteration* = 20, *param. Initialization Method* = *'Data Elements'*; *group Start Loc* = *1:blkLen:N*;
3. K-SVD dictionary training: [Ψ, output] = KSVD(x_i, param); the core is to use Equations (6)–(8) to generate a new dictionary Ψ through multiple iterative updates;
4. Sparse transformation: $x_i = \Psi\theta$;
5. Sensor matrix: $A_i = \Phi\Psi$;
6. Using the combination of Equations (1) and (2), the observation value of the data compression signal is obtained $y_i = A_i\theta$;
7. Signal transmission: The block-compressed signals $y_1, \ldots, y_i$ are successively transmitted through the data network;
8. **For** i = 1: size(x_i,2)/N (signal reconstruction);
9. θ_i = BSBL_BO (A_i, y_i, *groupStartLoc*, 0, *'prune_gamma'*, −1, *'max_iters'*, 20); the core is to use Formula (3)–(5) to solve the reconstructed signal θ_i;
10. Perform inverse sparse transformation to obtain reconstructed signal blocks: $x_1, \ldots, x_i$;
11. Connect the reconstructed signal blocks one by one to finally form a complete reconstructed signal: $x' = x_1 + x_2, + \ldots + x_i$;
12. **End**;
13. Output: x'.

3. Comprehensive Evaluation Index of Vibration Data Compression Effect

Reciprocating machinery vibration signal components are complex when compared to rotating machinery vibration signal components, noise pollution is severe, and a considerable amount of redundant data is created. There are numerous techniques in extant research to solve the data compression challenge. However, innumerable metrics are necessary to evaluate the data compression effect and performance benefits thoroughly. Although data compression technologies are widely utilized in the voice and image sectors, no standardized complete evaluation approach exists. As a result, while researching the vibration data compression method used in reciprocating equipment, it is vital to define a standard for evaluating the data compression effect. The following thorough assessment index of the data compression effect is produced by combining the structural properties of reciprocating equipment vibration data.

3.1. Data Compression Rate Evaluation Index

Data compression rate refers to the ratio of compressed data to the original data. It is a straightforward, intuitive, and easy-to-understand key indicator. Use CR (compressing ratio, CR) to represent the data compression ratio, and the range is set to (0, 1); then, the compression ratio is defined as follows [34]:

$$\mathrm{CR} = \frac{\mathrm{N} - \mathrm{M}}{\mathrm{N}} \tag{9}$$

N represents the original signal in the above equation, and M represents the compressed signal. The larger the CR value, the higher the data compression rate. When the data compression rate is higher, it does not mean that the data compression and reconstruction effect is better. It needs to be combined with the standard mean square error index for comprehensive evaluation.

3.2. Standard Mean Square Error Evaluation Index

The data compression rate is used to evaluate the ability of data compression. It shows that the loss rate of the original signal in the data compression process is very high. The accuracy of the reconstructed original signal is closely related to the compression rate. When the data compression rate is more significant, we cannot accurately restore the original signal reconstructed from the compressed signal. Therefore, based on data compression, MSE (mean square error, MSE) is used to represent the standard mean square error index, which the following equation can calculate [35]:

$$\mathrm{MSE} = \frac{\left\| Z' - Z \right\|_2}{\left\| Z \right\|_2} \tag{10}$$

Z represents the original signal in the equation above, and Z′ represents the reconstructed signal. The smaller the MSE value, the higher the accuracy of the data compression reconstructed signal. When the data compression rate is more significant, the MSE value is smaller, indicating better data compression and reconstruction effect.

3.3. Peak Signal-to-Noise Ratio Evaluation Index

The peak signal-to-noise ratio refers to the ratio of the original signal to the data com-pressed and reconstructed signal. In data compression, the loss of data information is reduced, and the quality of retaining the original data is improved as much as possible. PSNR (peak signal-to-noise ratio, PSNR) is used to express the peak signal-to-noise ratio [35], which the following equation can calculate:

$$\mathrm{PSNR} = 101g(z_{max}^2/(\frac{1}{\mathrm{N}}\sum_{j=1}^{N}(z_j - z_j')^2)) \tag{11}$$

z represents the original signal in the equation above, z′ represents the reconstructed signal, and z_{max} represents the maximum component. The greater the PSNR value, the higher the accuracy of the data compression and reconstruction signal, the closer it is to the original signal. It shows that the data compression and reconstruction effect is better.

3.4. Pearson Correlation Coefficient Evaluation Index

In evaluating the effect of data compression and reconstruction of the signal and using the two indicators of MSE and PSNR, usually, the Pearson correlation coefficient can also be used to evaluate the degree of correlation between the reconstructed signal and the original signal. Use r to represent the Pearson correlation coefficient, and the range is set to (−1,1), which can be calculated by the following equation [36]:

$$r_{z,z'} = \frac{N\sum ZZ' - \sum Z\sum Z'}{\sqrt{N\sum z^2 - (\sum z')^2}\sqrt{N\sum (z')^2 - (\sum z')^2}} \tag{12}$$

Z represents the original signal in the equation above, and Z′ represents the reconstructed signal. When the value of r is closer to 1, the similarity between the compressed and reconstructed signal and the original signal is higher and, conversely, the lower the similarity to the actual movement.

3.5. Comprehensive Evaluation Index in Time Domain

In fault prediction and health management, extracting characteristic parameters from vibration signals is crucial. Provide input conditions for further relevant analysis. For compressed data, the compression reconstruction algorithm should be able to recover from the compressed reconstructed signal similar to the original signal. Furthermore, in theory, it is identical to the actual feature parameters. Commonly used time-domain characteristic parameters mainly include mean value, root mean square value, variance and peak value, and other 12 indicators [37]. Under the same compression ratio, the feature parameters extracted from the reconstructed signal from compressed data are closer to the feature parameters extracted from the original signal, indicating that the less loss in the data compression process, the better the data restoration effect.

To better reflect the compression effect of the reconstructed signal in the time domain signal, the time domain characteristic index TT_i is defined, which can be calculated by the following equation:

$$TT_i = \left| \frac{T_i - \widetilde{T}_i}{T_i} \right|, \ i \in \{1, 2, 3, \cdots, 12\} \tag{13}$$

In the equation, T_i represents the time-domain feature value of the original signal, and $\widetilde{T}_i$ represents the time-domain feature value of the reconstructed signal. The smaller the TT_i value, the closer the time-domain characteristic index of the reconstructed signal and the original signal and the more accurate the data compression effect and restoration effect.

Similarly, in order to better evaluate the data compression effect of different compression algorithms, the comprehensive evaluation index KPI_t of time domain characteristics is defined, which can be calculated by the following equation:

$$KPI_t = \sum_{i=1}^{12} \omega_i \cdot TT_i \tag{14}$$

In the equation, ω_i represents the weight coefficient, which satisfies $\omega_i > 0$, and $\sum_{i=1}^{12} \omega_i = 1$. If there is no special case, the value is set to $\omega_i = 1/12$, i = 1,2,3, . . . ,12. The smaller the KPI_t value is, the closer the reconstructed signal data recovery is to the time domain index of the original signal, and the more accurate the corresponding data compression effect is.

4. Experimental Data Verification

4.1. Experiment Preparation

Figure 2 is the in-line six-cylinder diesel engine test bench used in the research. The test bench comprises three parts: diesel engine condition monitoring panel, diesel engine, and vibration signal data acquisition system. The diesel engine status monitoring panel can control the ignition, acceleration, and flameout of the diesel engine. The instrument reflects the engine speed, water temperature, voltage, and remaining oil. Preset 6 intake valve clearance state modes under different working conditions include one normal status and five other fault states. The detailed parameters of the dataset are shown in Table 1. To obtain valid data samples, four vibration sensors are arranged on the cylinder head of the diesel engine, as shown in Figure 2b. Among them, the sampling frequency of data acquisition is set to 20 kHz, and the duration of each acquisition is set to 10 s. Each failure mode collects ten sets of data samples, and each data group contains 200,000 points (20 kHZ sampling for 10 s), as shown in Figure 3.

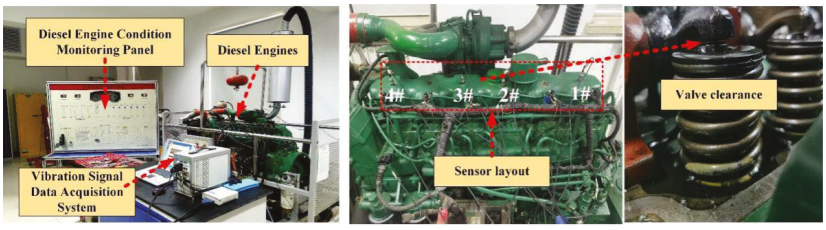


Figure 2. Diesel engine preset failure experiment environment: (**a**) test bench; (**b**) intake valve clearance failure.

Table 1. Experimental Dataset of Valve Clearance in Different Working Conditions.

No.	Dataset	State	Rotating Speed	Inlet Valve Clearance
1	Valve_800_3mm	Normal Status	800	0.3 mm
2	Valve_800_7mm	Fault 1	800	0.7 mm
3	Valve_1200_3mm	Fault 2	1200	0.3 mm
4	Valve_1200_4mm	Fault 3	1200	0.4 mm
5	Valve_1200_5mm	Fault 4	1200	0.5 mm
6	Valve_1200_7mm	Fault 5	1200	0.7 mm

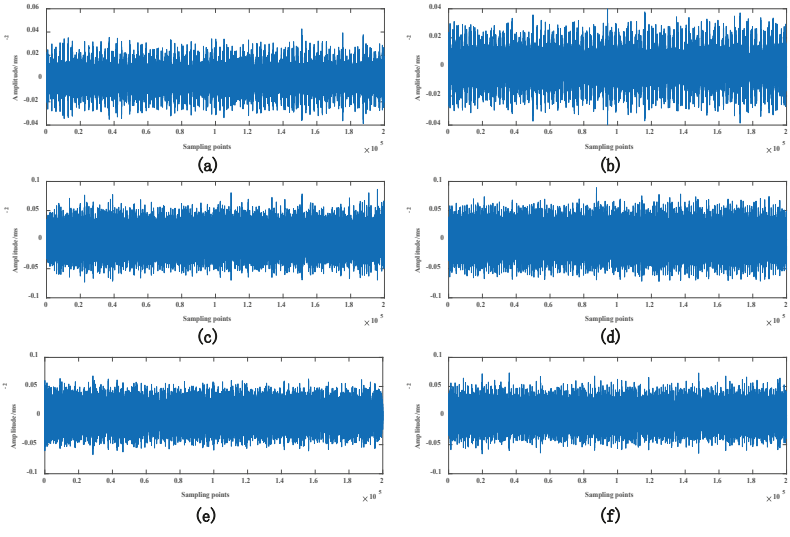


Figure 3. Experimental dataset: (**a**) Valve_800_3mm; (**b**) Valve_800_7mm; (**c**) Valve_1200_3mm; (**d**) Valve_1200_4mm; (**e**) Valve_1200_5mm; and (**f**) Valve_1200_7mm.

4.2. Comparison of BSBL-BO Algorithm with Other Compression and Reconstruction Algorithms

4.2.1. Evaluation Index of Reconstructed Signal MSE under the Same Compression Ratio

Compare and analyze BSBL-BO algorithm with block sparse Bayesian learning-expectation-maximization (BSBL-EM), compressive sampling matched pursuit (CoSaMP), BP, OMP, and ROMP algorithm. Use the Valve_1200_7mm dataset to verify and analyze the reconstruction algorithm, as shown in Figure 4. To ensure the reconstruction performance of the algorithm, the data compression rate is uniformly set to 0.5, and the sparse dictionary matrix uniformly uses the K-SVD generation method. From the analysis results in Figure 4, it can be seen that the smaller the MSE index, the higher the reconstruction accuracy, indicating that under the

same parameter setting conditions, the proposed BSBL-BO reconstruction algorithm has more advantages than its reconstruction algorithm. The recovered reconstructed signal is closer to the original signal, proving that the data compression and reconstruction effect is better.

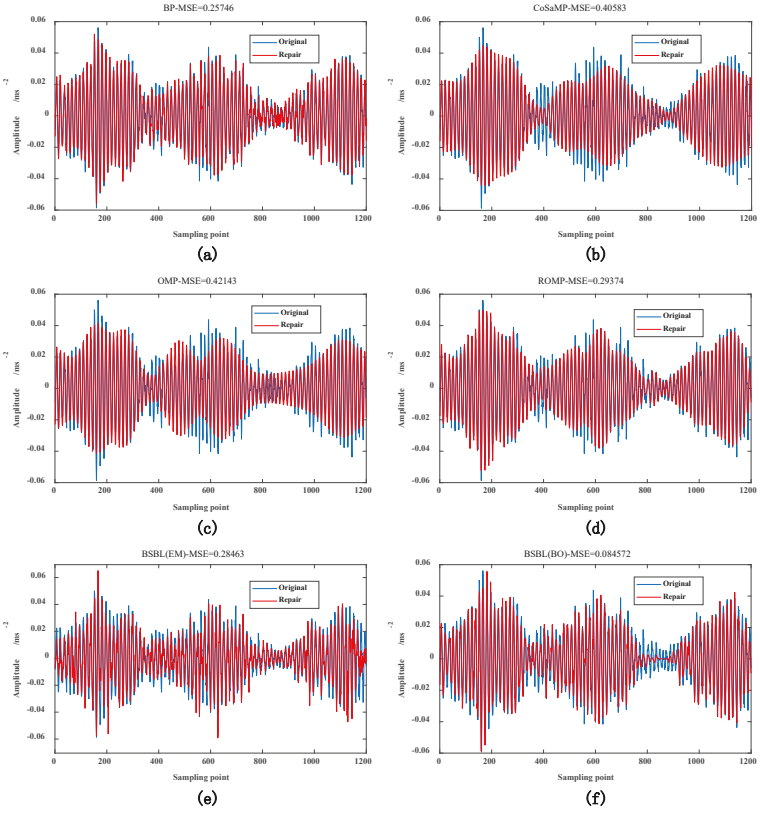


Figure 4. Valve_1200_7mm dataset: (**a**) BP algorithm (MSE = 0.25746); (**b**) CoSaMP algorithm (MSE = 0.40583); (**c**) OMP algorithm (MSE = 0.42143); (**d**) ROMP algorithm (MSE = 0.29374); (**e**) BSBL(EM) algorithm (MSE = 0.28463); and (**f**) BSBL(BO) algorithm (MSE = 0.084572).

4.2.2. Evaluation Index of Reconstructed Signal MSE under Different Compression Ratios

A comprehensive analysis of the algorithm's influence on different compression ratio changes is carried out. Six different datasets are used to verify the compression and reconstruction algorithm. Among them, each dataset sets 13 kinds of compression ratios. Each compression rate is performed 100 times of MSE calculation. Find the corresponding variance σ and average μ, and use the 95% confidence interval $(\mu - 2\sigma, \mu + 2\sigma)$ method to express, as shown in Figure 5. It can be seen from the analysis result of Figure 5, when CR < 0.6, the MSE index of the BSBL-BO reconstruction algorithm proposed in this paper is smaller than other reconstruction algorithms. Know the accuracy, superiority, and effectiveness of the proposed method. When CR > 0.6, all reconstruction algorithms have a more considerable MSE value as the compression ratio increases. It means that the data lose essential information during the compression process, resulting in a significant reduction in the reconstruction accuracy. The ROMP algorithm has the most considerable MSE value and the lowest reconstruction accuracy. As the compression ratio increases, the reconstruction accuracy also decreases. Conversely, the lower the compression ratio, the higher the reconstruction accuracy. Therefore, after being verified by six different datasets, under the premise of ensuring a specific data compression rate and sure re-construction accuracy,

when the CR = 0.5, it is confirmed that the method proposed in this paper is the best for data compression of vibration signals.

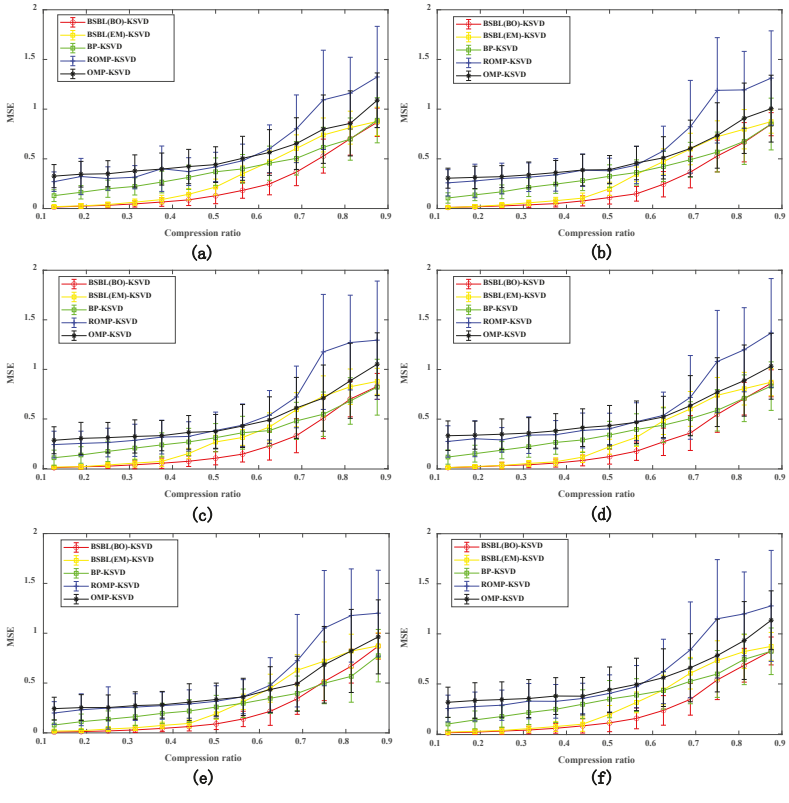
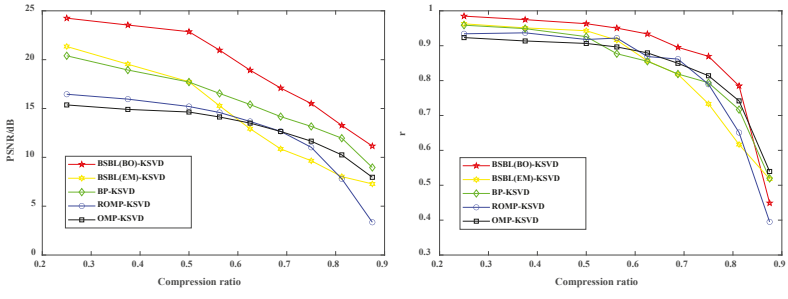


Figure 5. Comparative analysis of MSE 95% confidence intervals of six datasets: (**a**) Valve_800_3mm; (**b**) Valve_800_7mm; (**c**) Valve_1200_3mm; (**d**) Valve_1200_4mm; (**e**) Valve_1200_5mm; and (**f**) Valve_1200_7mm.

4.2.3. Peak Signal-to-Noise Ratio Evaluation and Pearson Correlation Coefficient under Different Compression Ratios

As shown in Figure 6a, the method proposed in this paper has significant advantages compared with other methods. When the compression ratio increases, the PSNR value decreases, indicating that more data information is lost during data compression. The reconstructed signal is different from the original signal and has a low peak signal-to-noise ratio. Combining them with the MSE metric is recommended when evaluating data compression results. The larger the PSNR index, the smaller the MSE index and the better the data compression effect. As shown in Figure 6b, The BSBL-KSVD method also outperforms other ways and the Pearson correlation coefficient increases as the compression ratio decreases. The results show that much of the original signal's information is preserved in the data when compressed. Therefore, the reconstructed signal has a high similarity with the original signal. When evaluating the effect of data compression, it is recommended to combine the MSE indicator. The smaller the MSE index, the higher the Pearson correlation coefficient, and the better the data compression effect. From a comprehensive analysis, when CR = 0.5, it is proven that the method proposed in this paper has the best compression effect and is more suitable for data compression of vibration signals.



(a) (b)

Figure 6. Comparative analysis of different compression and reconstruction methods: (**a**) peak signal-to-noise ratio evaluation and (**b**) Pearson correlation coefficient.

4.2.4. Comprehensive Evaluation Index of Reconstructed Signal in Time Domain under Different Compression Ratios

Next, to better evaluate the pros and cons of the recovered reconstructed signal, Using the same compression and reconstruction algorithm and data in Section 4.2.1 and combined with the time domain comprehensive evaluation index KPI_t for comparative analysis, the KPI_t index weights are all set to 1/12, and the analysis results are shown in Figure 7. It can be seen from Figure 7 that the smaller the KPI_t index is, it means that the restored reconstructed signal retains most of the original signal. The time-domain characteristics of the reconstructed signal are closer to the frequency domain characteristics of the original signal, which proves that the proposed method has the best data compression effect. In a comprehensive analysis, the corresponding KPI_t index is more minor when the compression rate is lower, indicating that the data compression effect is better. Therefore, it is proven that when the compression ratio CR = 0.5, the compression effect of the BSBL-KSVD algorithm proposed in this paper is optimal, which is more suitable for data compression.

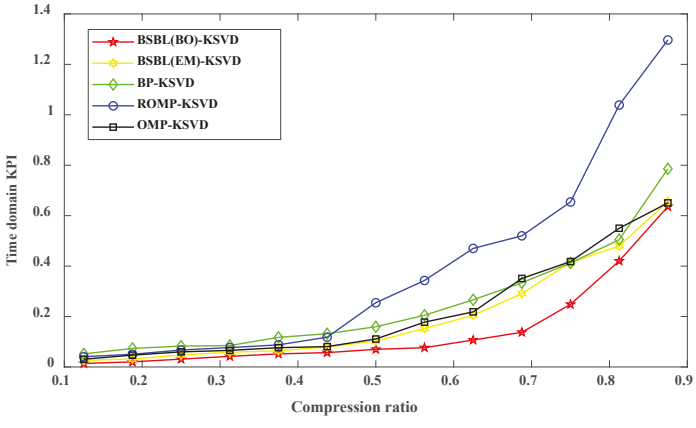


Figure 7. Time-domain comprehensive evaluation index of reconstructed signal under different compression ratios.

4.3. K-SVD Dictionary and Other Dictionary Effect Verification Comparison

In data compression, the sparse representation of the signal is critical since the sparse representation of the static dictionary has relatively low complexity. Assuming that the signal feature information is consistent with the atomic data in the dictionary, a more accurate and effective sparse representation can be obtained. Commonly used classic fixed dictionaries to obtain the sparse dictionary matrix include DFT, DWT, DCT, etc.

Therefore, the K-SVD dictionary is compared and analyzed with the DCT, DFT, and DWT dictionaries. Use the Valve_1200_7mm dataset to verify and scrutinize the reconstruction algorithm. First, the 200,000 sampling points of the original signal only select the first 64,000 sampling points for block compression. The length of each signal block is set to 80 sampling points, which are divided into 800 blocks. A Gaussian random matrix uniformly generates the observation matrix. Secondly, for the K-SVD dictionary, the number of atoms is set to 50, the number of iterations is set to 20, and 300 blocks of signals are trained each time. The remaining 500 pieces of signs are used to verify the validity of the dictionary. Finally, the single variable principle is adopted, and the BSBL-BO compression and reconstruction algorithm is uniformly adopted. It is applied to different sparse dictionaries and verified from the MSE evaluation index, peak signal-to-noise ratio evaluation index, and Pearson correlation coefficient evaluation index.

4.3.1. Evaluation Index of Reconstructed Signal MSE under Different Compression Ratios

As shown in Figure 8a, the compression effect of vibration data based on the K-SVD dictionary is better than that of other dictionaries. The blue lines represents the original signal, and the red lines represents the reconstructed signal in Figure 8b. Observing Figure 8b, we can find that when CR = 0.5, the waveform of the reconstructed signal based on the K-SVD dictionary is closer to the original signal than in other dictionaries. When CR > 0.7, the greater the MSE index, and the data reconstruction effect is worse. Therefore, it is proven that the proposed method is more suitable for data compression when CR = 0.5.

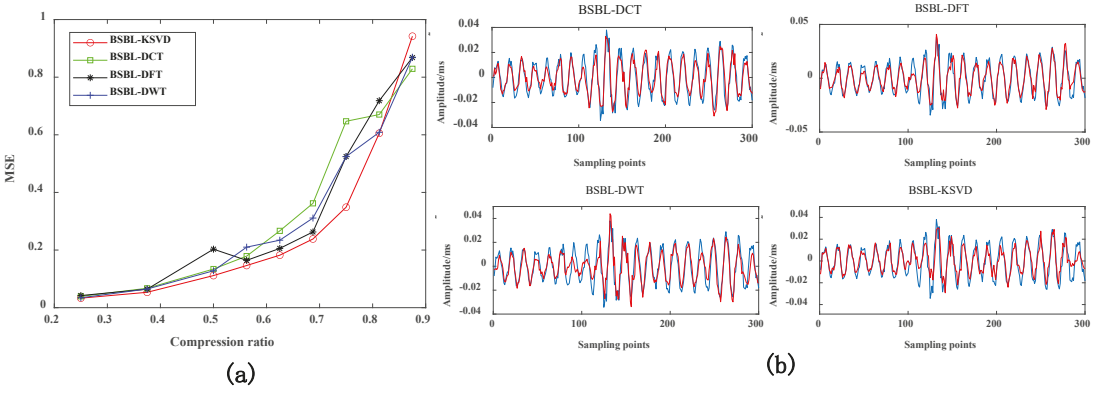


Figure 8. (**a**) Comparison of MSE indicators of different sparse dictionaries; (**b**) when CR = 0.5, the reconstruction signal comparison of different dictionaries.

4.3.2. Peak Signal-to-Noise Ratio of Reconstructed Signal under Different Compression Ratios

In Figure 9b, the blue lines represents the original signal, and the red lines represents the reconstructed signal. As can be seen from Figure 9, The data compression effect based on the K-SVD dictionary is better than other dictionaries. When CR > 0.7, the PSNR indicator becomes smaller as the compression rate increases. It shows that a great deal of data information is lost in data compression. The recovered reconstructed signal is quite different from the original signal, and the peak signal-to-noise ratio will naturally become smaller. It needs to be evaluated in combination with MSE indicators. When the MSE index of the reconstructed signal is smaller, and the PSNR index is more extensive, it proves that the performance of the proposed method is better. Therefore, when CR = 0.5, it is more suitable for data compression.

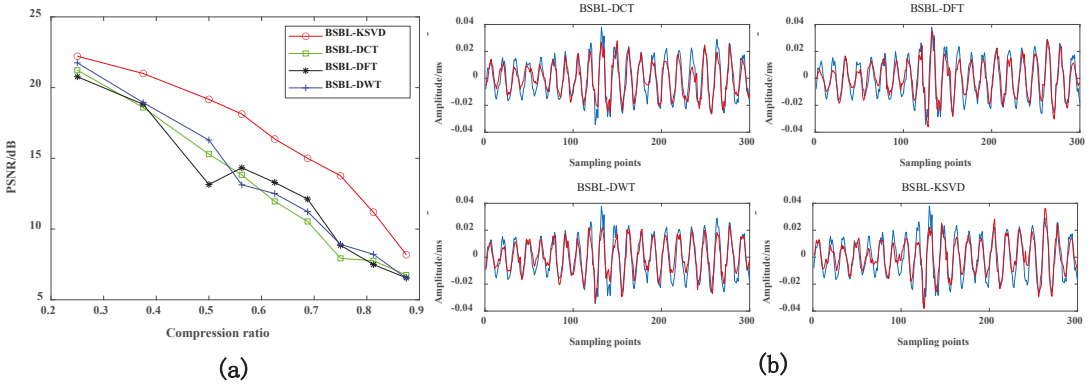


Figure 9. (**a**) Comparison of PSNR indicators for different sparse dictionaries; (**b**) when CR = 0.7, the reconstruction signal comparison of different dictionaries.

4.3.3. Pearson Correlation Coefficient of Reconstructed Signal under Different Compression Ratios

In Figure 10b, the blue lines represents the original signal, and the red lines represents the reconstructed signal. As can be seen from Figure 10, The data compression effect based on the K-SVD dictionary is also better than other dictionaries. When CR < 0.6, as the compression rate gradually decreases, the more significant the Pearson correlation coefficient, and the data retains a large amount of original signal information during the compression process. The similarity between the restored reconstructed signal and the original signal becomes higher. Therefore, it needs to be used in conjunction with the MSE indicator. When the MSE indicator is more minor, and the Pearson correlation coefficient is more significant, it is proven that the compression effect of this method is the best. When CR = 0.5, it is more suitable for data compression.

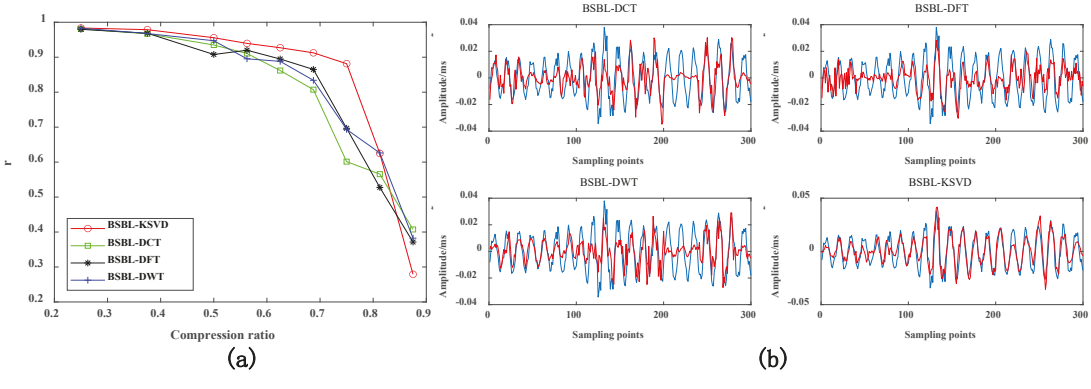


Figure 10. (**a**) Comparison of r indicators of different sparse dictionaries; (**b**) when CR = 0.6, the reconstruction signal comparison of different dictionaries.

5. Application of Compressed and Reconstructed Signal in Fault Diagnosis

To further verify the effectiveness of the BSBL-KSVD compression reconstruction method proposed in this paper in fault diagnosis, two forms of fault classification accuracy are adopted: naive Bayes classifier (NBC) and support vector machines (SVM). A comprehensive evaluation is performed to check the quality of the compressed and reconstructed signal. The higher the classification accuracy, the closer the reconstructed signal is to the original signal. The fault test dataset in Table 1 is used for fault diagnosis, and ten sets of samples are taken for each fault

state. There are 200,000 sampling points in each group, with 5500 sampling points as a group, divided into 360 groups of samples and six failure states, a total of 6 × 360 = 2160 samples. Each sensor's fault state is extracted from the time and frequency domains, including 22 characteristic parameters such as mean values, root mean square values, variance, and peak values [37,38]. Each sensor forms a 22 × 2160 feature matrix.

5.1. Comparative Analysis of Fault Classification under Different Compression Ratios

Therefore, select sensor 1–4# data to form a feature matrix of 88 × 2160. After dimensionality reduction by the stacked sparse autoencoder (SSAE) method, SSAE input nodes are set to 88, and the hidden layer parameters are 50 and 22, respectively. The sparsity ratio is set to 0.1, the weight adjustment coefficient is set to 0.000002, and the sparsity penalty weight is set to 0.0002. A new 22 × 2160 feature matrix is obtained, divided into 1800 training samples and 360 test samples. Using the built-in classification learning tool of Matlab 2020. Among them, 1800 training samples adopt the K-fold cross-validation method and take K = 10. Input to the classifier method: NBC and SVM are trained, and the optimal training model is obtained. Then, input 360 test samples into the trained model for fault identification. Obtain the results of fault diagnosis accuracy, as shown in Table 2. The confusion matrix of the fault diagnosis results is shown in Figure 11. It can be seen from the effects that the higher the compression rate CR value, the lower the fault classification accuracy rate. SVM has an accuracy rate of 96.39% for the original signal fault diagnosis, while NBC has an accuracy rate of 90.83%. When CR = 0.25, the classification accuracy of SVM reaches 95.56%, while that of NBC is 89.72%. It is very close to the classification result of the original signal. We obtain the same conclusion as in Section 4.2: The BSBL-KSVD compression reconstruction method is suitable for high data compression.

Table 2. Comparative analysis of fault classification under different compression ratios.

State	Original Signal		CR = 0.25		CR = 0.5		CR = 0.75	
	SVM	**NBC**	**SVM**	**NBC**	**SVM**	**NBC**	**SVM**	**NBC**
Normal Status	96.67%	91.67%	95.00%	91.67%	93.33%	96.67%	85.00%	88.33%
Fault 1	95.00%	90.00%	91.67%	88.33%	88.33%	81.67%	86.67%	78.33%
Fault 2	93.33%	86.67%	93.33%	78.33%	86.67%	75.00%	88.33%	80.00%
Fault 3	93.33%	88.33%	95.00%	86.67%	93.33%	86.67%	98.33%	91.67%
Fault 4	100.0%	95.00%	100.0%	95.00%	100.0%	95.00%	98.33%	86.67%
Fault 5	100.0%	93.33%	98.33%	98.33%	96.67%	96.67%	100.0%	96.67%
Total Accuracy	**96.39%**	**90.83%**	**95.56%**	**89.72%**	**93.06%**	**88.61%**	**92.78%**	**86.95%**

As shown in Figure 11, whether it is the original signal or different compressed signals, the fault recognition rates for fault 1, fault 2, and fault 3 are relatively low. Among them, the classification result of defect one increases with the increase of compression rate, while the accuracy rate gradually decreases. Therefore, the BSBL-KSVD compression reconstruction method proposed in this paper hopes to find the optimal balance between the fault diagnosis accuracy and the wireless network transmission. It shows that this kind of fault signal contains fewer fault features, which increases the difficulty of fault classification. It can be recognized if the fault diagnosis accuracy rate is more than 90%. Then, when CR = 0.5, the compressed vibration signal during wireless transmission will significantly reduce the constraint of network bandwidth and improve the transmission efficiency.

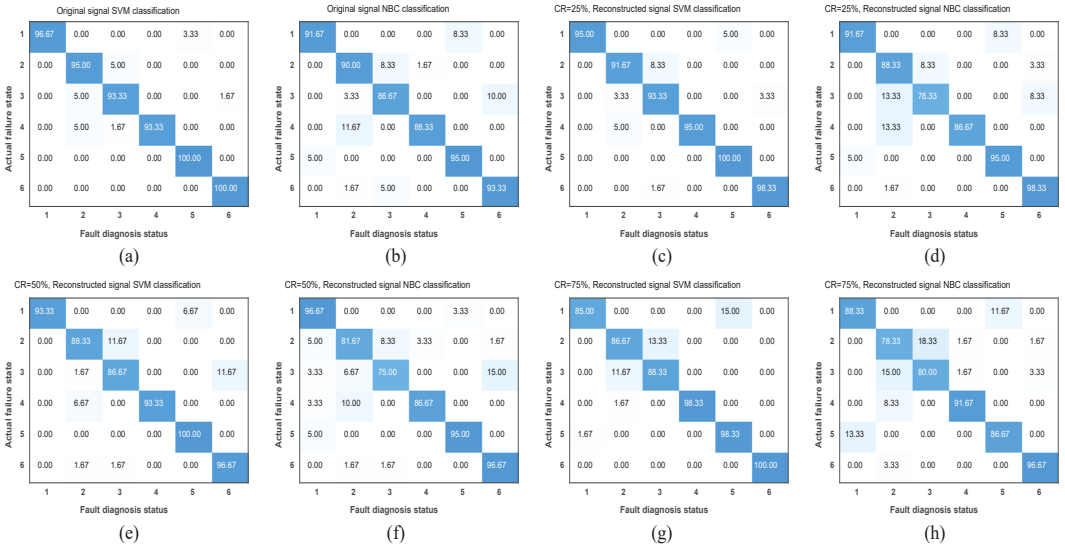


Figure 11. Comparative analysis of fault classification under different compression ratios: (**a**) Original signal SVM classification; (**b**) original signal NBC classification; (**c**) CR = 0.25, SVM classification; (**d**) CR = 0.25, NBC classification; (**e**) CR = 0.5, SVM classification; (**f**) CR = 0.5, NBC classification; (**g**) CR = 0.75, SVM classification; and (**h**) CR = 0.75, NBC classification.

5.2. Comparative Analysis of Fault Diagnosis Results of Different Compression and Reconstruction Methods

We compare and analyze BSBL-KSVD with other compression and reconstruction algorithms and use the experimental data in Section 4.1 to verify the method's effectiveness. First, use five compression and reconstruction algorithms to process the original data in Table 1 with three compression ratios (i.e., CR = 0.25, CR = 0.50, CR = 0.75). Then, using the feature extraction method in Section 5, different fault feature matrices of 88 × 2160 are extracted from the other reconstructed signals of the four sensors. The SSAE method is also used for dimensionality reduction where the SSAE parameter settings are the same as in Section 5.1. Finally, for the three compression ratios under each compression and reconstruction method. We can obtain a new 22 × 2160 feature matrix after dimension reduction, respectively, and divide it into 1800 training samples and 360 testing samples. In addition, use the built-in SVM classification tool of Matlab 2020 for fault diagnosis. The relevant parameter settings are the same as those in Section 5.1, and the final diagnosis results under different compression ratios are shown in Tables 3–5.

Table 3. When CR = 0.25, the comparative analysis of fault classification of different compression and reconstruction methods.

State	BSBL(BO)-KSVD	BSBL(EM)-KSVD	BP-KSVD	ROMP-KSVD	OMP-KSVD
Normal Status	95.00%	91.33%	83.33%	86.67%	85.33%
Fault 1	91.67%	85.00%	95.00%	86.67%	90.00%
Fault 2	93.33%	83.33%	81.67%	83.33%	82.67%
Fault 3	95.00%	93.33%	88.33%	81.67%	80.67%
Fault 4	100.0%	95.67%	85.00%	95.00%	90.33%
Fault 5	98.33%	100.0%	95.00%	91.33%	92.00%
Total Accuracy	**95.56%**	**91.44%**	**88.06%**	**87.45%**	**86.83%**

Table 4. When CR = 0.5, the comparative analysis of fault classification of different compression and reconstruction methods.

State	BSBL(BO)-KSVD	BSBL(EM)-KSVD	BP-KSVD	ROMP-KSVD	OMP-KSVD
Normal Status	93.33%	87.67%	78.33%	80.67%	83.67%
Fault 1	88.33%	83.33%	80.00%	82.33%	72.00%
Fault 2	86.67%	85.00%	75.00%	70.33%	72.33%
Fault 3	93.33%	88.33%	81.67%	76.00%	68.00%
Fault 4	100.0%	95.00%	92.33%	87.67%	82.33%
Fault 5	96.67%	90.67%	88.33%	88.00%	85.67%
Total Accuracy	**93.06%**	**88.33%**	**82.61%**	**80.83%**	**77.33%**

Table 5. When CR = 0.75, the comparative analysis of fault classification of different compression and reconstruction methods.

State	BSBL(BO)-KSVD	BSBL(EM)-KSVD	BP-KSVD	ROMP-KSVD	OMP-KSVD
Normal Status	85.00%	81.00%	72.67%	76.67%	65.00%
Fault 1	86.67%	77.33%	75.00%	65.00%	74.67%
Fault 2	88.33%	77.33%	71.33%	73.67%	67.33%
Fault 3	98.33%	89.00%	83.33%	67.33%	67.33%
Fault 4	98.33%	82.67%	75.67%	70.00%	72.00%
Fault 5	100.0%	85.33%	80.00%	75.67%	63.67%
Total Accuracy	**92.78%**	**82.11%**	**76.33%**	**71.39%**	**68.33%**

From the diagnostic results in Table 3, it can be seen that when CR = 0.25, the diagnostic results of the BSBL-KSVD method are better than other compression and reconstruction methods. The diagnostic results of BP, ROMP, and OMP algorithms are less than 90%, indicating that the reconstruction accuracy of these three types of strategies is not high. Some critical data information is lost during data compression.

From the analysis of the diagnostic results in Table 4, when CR = 0.5, only the diagnostic results of the BSBL-KSVD method are > 90%, The diagnostic results of the other four compression methods were lower than 90%, and the lowest diagnostic result of the OMP method was only 77.33%. It shows that with the increase of compression ratio, the diagnosis result gradually decreases.

From the results in Table 5, when CR = 0.75, the diagnostic results of BP, ROMP, and OMP methods are all below 80%, while the BSBL-KSVD method can reach more than 90%. Compared with other compression methods, the method proposed in this paper has good robustness and superiority.

To sum up, the diagnosis results of the BSBL-KSVD method are better than other compression and reconstruction methods under different compression ratios. In the case of weighing various pros and cons, it is assumed that the diagnostic result is >90% and has a high data compression rate. This is a good reference for applying data compression to mechanical fault diagnosis.

6. Conclusions

This paper proposes a method of compression and reconstruction of diesel engine vibration signal based on BSBL-KSVD, which is practical and feasible, and compared with other methods, there are advantages. To effectively verify the pros and cons of the BSBL-KSVD algorithm proposed in this study regarding data compression effects, use the CR indicator, MSE indicator, PSNR indicator, r indicator, and KPIs indicator for verification and, finally, compressed and reconstructed signals for fault diagnosis case analysis. The experimental results show that the compression effect of the BSBL-KSVD algorithm is optimal when the compression rate CR = 0.5. The recovered reconstructed signal is closer to the original signal, and good classification accuracy is obtained, which has a good engineering application prospect.

Although the proposed method has achieved good results, we can still improve it in the following aspects: First, this research did not focus on using the reconstructed signal to perform signal repair and noise reduction preprocessing in the follow-up. We will conduct detailed research using methods such as double sparse dictionary learning; second, it did not consider integrating the algorithm with the data acquisition hardware. In the subsequent investigation, embedding the algorithm into FPGA improves front-end data acquisition, transmission performance, and efficiency.

Author Contributions: Data curation, H.B. and Y.M.; Resources, L.W. and Y.M.; Supervision, X.J.; Validation, L.W. and X.J.; Writing—original draft, H.B.; Writing—review & editing, H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be provided upon request.

Conflicts of Interest: Authors declare that they have no conflict of interest.

References

1. Wang, R.; Chen, H.; Guan, C. Random convolutional neural network structure: An intelligent health monitoring scheme for diesel engines. *Measurement* **2021**, *171*, 108786. [CrossRef]
2. Lan, F.; Jiang, Y.; Wang, H. Performance Prediction Method of Prognostics and Health Management of Marine Diesel Engine. *J. Phys. Conf. Ser.* **2020**, *1670*, 12014. [CrossRef]
3. Jin, C.; Zhao, W.; Liu, Z.; Lee, J.; He, X. A vibration-based approach for diesel engine fault diagnosis. In Proceedings of the 2014 International Conference on Prognostics and Health Management, Cheney, WA, USA, 22–25 June 2014; pp. 1–9.
4. Wang, X.; Wang, Y.; Shi, X.; Gao, L.; Li, P. A probabilistic multimodal optimization algorithm based on Buffon principle and Nyquist sampling theorem for noisy environment. *Appl. Soft Comput.* **2021**, *104*, 107068. [CrossRef]
5. Antonopoulos, C.P.; Voros, N.S. A data compression hardware accelerator enabling long-term biosignal monitoring based on ultra-low power IoT platforms. *Electronics* **2017**, *6*, 54. [CrossRef]
6. Ma, N. Distributed video coding scheme of multimedia data compression algorithm for wireless sensor networks. *EURASIP J. Wirel. Commun. Netw.* **2019**, *2019*, 254. [CrossRef]
7. Yi, J.M.; Oh, E.J.; Noh, D.K.; Yoon, I. Energy-aware data compression and transmission range control for energy-harvesting wireless sensor networks. *Int. J. Distrib. Sens. Netw.* **2017**, *13*, 1550147717705785. [CrossRef]
8. Hameed, M.E.; Ibrahim, M.M.; Abd Manap, N.; Mohammed, A.A. A lossless compression and encryption mechanism for remote monitoring of ECG data using Huffman coding and CBC-AES. *Future Gener. Comput. Syst.* **2020**, *111*, 829–840. [CrossRef]
9. Khosravy, M.; Cabral, T.W.; Luiz, M.M.; Gupta, N.; Crespo, R.G. Random Acquisition in Compressive Sensing: A Comprehensive Overview. *Int. J. Ambient. Comput. Intell.* **2021**, *12*, 140–165. [CrossRef]
10. Li, Y.; Zheng, F.; Xiong, Q.; Zhang, W. A secondary selection-based orthogonal matching pursuit method for rolling element bearing diagnosis. *Measurement* **2021**, *176*, 109199. [CrossRef]
11. Rahim, T.; Novamizanti, L.; Ramatryana IN, A.; Shin, S.Y. Compressed medical imaging based on average sparsity model and reweighted analysis of multiple basis pursuit. *Comput. Med. Imaging Graph.* **2021**, *90*, 101927. [CrossRef]
12. Dey, M.R.; Shiraz, A.; Sharif, S.; Lota, J.; Demosthenous, A. Dictionary selection for compressed sensing of EEG signals using sparse binary matrix and spatiotemporal sparse Bayesian learning. *Biomed. Phys. Eng. Express* **2020**, *6*, 65024. [CrossRef] [PubMed]
13. Liu, R.; Shu, M.; Chen, C. ECG signal denoising and reconstruction based on basis pursuit. *Appl. Sci.* **2021**, *11*, 1591. [CrossRef]
14. Cheng, C.; Lin, D. Based on Compressed Sensing of Orthogonal Matching Pursuit Algorithm Image Recovery. *J. Internet Things* **2020**, *2*, 37–45. [CrossRef]
15. Sajjad, M.; Mehmood, I.; Baik, S.W. Sparse coded image super-resolution using K-SVD trained dictionary based on regularized orthogonal matching pursuit. *Bio-Med. Mater. Eng.* **2015**, *26*, S1399–S1407. [CrossRef] [PubMed]
16. Zhang, Z.; Rao, B.D. Extension of SBL Algorithms for the Recovery of Block Sparse Signals with Intra-Block Correlation. *IEEE Trans. Signal Process.* **2013**, *61*, 2009–2015. [CrossRef]
17. Mahrous, H.; Ward, R. Block sparse compressed sensing of electroencephalogram (EEG) signals by exploiting linear and non-linear dependencies. *Sensors* **2016**, *16*, 201. [CrossRef] [PubMed]
18. Li, G.; Ye, W.; Lao, G.; Kong, S.; Yan, D. Narrowband Interference Separation for Synthetic Aperture Radar via Sensing Matrix Op-timization-Based Block Sparse Bayesian Learning. *Electronics* **2019**, *8*, 458. [CrossRef]
19. Almasri, N.; Sadhu, A.; Chaudhuri, S.R. Toward compressed sensing of structural monitoring data using discrete cosine transform. *J. Comput. Civ. Eng.* **2020**, *34*, 4019041. [CrossRef]
20. Borghesani, P.; Pennacchi, P.; Chatterton, S.; Ricci, R. The velocity synchronous discrete Fourier transform for order tracking in the field of rotating machinery. *Mech. Syst. Signal Process.* **2014**, *44*, 118–133. [CrossRef]
21. Belkacemi, B.; Saad, S.; Ghemari, Z.; Zaamouche, F.; Khazzane, A. Detection of induction motor improper bearing lubrication by discrete wavelet transforms (DWT) decomposition. *Instrum. Mes. Métrologie* **2020**, *19*, 347–354. [CrossRef]

22. Liang, K.; Zhao, M.; Lin, J.; Jiao, J. An information-based K-singular-value decomposition method for rolling element bearing diagnosis. *ISA Trans.* **2020**, *96*, 444–456. [CrossRef] [PubMed]
23. Hosseini, M.; Riahi, M.A. Using input-adaptive dictionaries trained by the method of optimal directions to estimate the permeability model of a reservoir. *J. Appl. Geophys.* **2019**, *165*, 16–28. [CrossRef]
24. Li, Q.; Liang, S.Y. Microstructure Images Restoration of Metallic Materials Based upon KSVD and Smoothing Penalty Sparse Representation Approach. *Materials* **2018**, *11*, 637. [CrossRef] [PubMed]
25. Yang, J.; Zhang, X.; Peng, W.; Liu, Z. A novel regularized K-SVD dictionary learning based medical image super-resolution algorithm. *Multimed. Tools Appl.* **2016**, *75*, 13107–13120. [CrossRef]
26. Gu, C.; Qiao, X.-Y.; Li, H.; Jin, Y. Misfire fault diagnosis method for diesel engine based on MEMD and dispersion entropy. *Shock Vib.* **2021**, *2021*, 9213697. [CrossRef]
27. Chen, K.; Mao, Z.; Zhao, H.; Jiang, Z.; Zhang, J. A variational stacked autoencoder with harmony search optimizer for valve train fault diagnosis of diesel engine. *Sensors* **2019**, *20*, 223. [CrossRef] [PubMed]
28. Wang, B.; Ke, H.; Ma, X.; Yu, B. Fault Diagnosis Method for Engine Control System Based on Probabilistic Neural Network and Support Vector Machine. *Appl. Sci.* **2019**, *9*, 4122. [CrossRef]
29. Zhang, X.; Hu, N.; Hu, L.; Chen, L. A bearing fault detection method with low-dimensional compressed measurements of vibration signal. *J. Vibroeng.* **2015**, *17*, 1253–1265.
30. Tang, G.; Hou, W.; Wang, H.; Luo, G.; Ma, J. Compressive sensing of roller bearing faults via harmonic detection from under-sampled vibration signals. *Sensors* **2015**, *15*, 25648–25662. [CrossRef]
31. Du, Z.; Chen, X.; Zhang, H.; Miao, H.; Guo, Y.; Yang, B. Feature identification with compressive measurements for machine fault diagnosis. *IEEE Trans. Instrum. Meas.* **2016**, *65*, 977–987. [CrossRef]
32. Zhang, X.; Ma, Y.; Gao, Y.; Zhang, W. Autonomous compressive-sensing-augmented spectrum sensing. *IEEE Trans. Veh. Technol.* **2018**, *67*, 6970–6980. [CrossRef]
33. Aharon, M.; Elad, M.; Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322. [CrossRef]
34. Šaliga, J.; Andráš, I.; Dolinský, P.; Michaeli, L.; Kováč, O.; Kromka, J. ECG compressed sensing method with high compression ratio and dynamic model reconstruction. *Measurement* **2021**, *183*, 109803. [CrossRef]
35. Dileep BP, V.; Das, T.; Dutta, P.K. Greedy algorithms for diffuse optical tomography reconstruction. *Opt. Commun.* **2018**, *410*, 164–173. [CrossRef]
36. Zhiyong, L.; Hongdong, Z.; Ruili, Z.; Kewen, X.; Qiang, G.; Yuhai, L. Fault identification method of diesel engine in light of pearson correlation coefficient diagram and orthogonal vibration signals. *Math. Probl. Eng.* **2019**, *2019*, 2837580. [CrossRef]
37. Samuel, P.D.; Pines, D.J. A review of vibration-based techniques for helicopter transmission diagnostics. *J. Sound Vib.* **2005**, *282*, 475–508. [CrossRef]
38. Liu, Z.; Qu, J.; Zuo, M.J.; Xu, H.B. Fault level diagnosis for planetary gearboxes using hybrid kernel feature selection and kernel Fisher discriminant analysis. *Int. J. Adv. Manuf. Technol.* **2013**, *67*, 1217–1230. [CrossRef]

sensors

MDPI

Article

A Novel Fault Diagnosis Method of Rolling Bearing Based on Integrated Vision Transformer Model

Xinyu Tang [1,2], Zengbing Xu [1,2,3,*] and Zhigang Wang [1,2]

[1] Key Laboratory of Metallurgical Equipment and Control Technology, Wuhan University of Science and Technology, Ministry of Education, Wuhan 430081, China; tangxinyu@wust.edu.cn (X.T.); wzhigang@wust.edu.cn (Z.W.)
[2] Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China
[3] The State Key Laboratory of Digital Manufacturing Equipment & Technology, Huazhong University of Science and Technology, Wuhan 430074, China
* Correspondence: xuzb@wust.edu.cn; Tel.: +86-027-68862252

Abstract: In order to improve the diagnosis accuracy and generalization of bearing faults, an integrated vision transformer (ViT) model based on wavelet transform and the soft voting method is proposed in this paper. Firstly, the discrete wavelet transform (DWT) was utilized to decompose the vibration signal into the subsignals in the different frequency bands, and then these different subsignals were transformed into a time–frequency representation (TFR) map by the continuous wavelet transform (CWT) method. Secondly, the TFR maps were input with respective to the multiple individual ViT models for preliminary diagnosis analysis. Finally, the final diagnosis decision was obtained by using the soft voting method to fuse all the preliminary diagnosis results. Through multifaceted diagnosis tests of rolling bearings on different datasets, the diagnosis results demonstrate that the proposed integrated ViT model based on the soft voting method can diagnose the different fault categories and fault severities of bearings accurately, and has a higher diagnostic accuracy and generalization ability by comparison analysis with integrated CNN and individual ViT.

Keywords: vision transformer; integrated vision transformer; fault diagnosis; rolling bearing

Citation: Tang, X.; Xu, Z.; Wang, Z. A Novel Fault Diagnosis Method of Rolling Bearing Based on Integrated Vision Transformer Model. *Sensors* **2022**, *22*, 3878. https://doi.org/10.3390/s22103878

Academic Editors: Shilong Sun, Changqing Shen and Dong Wang

Received: 20 April 2022
Accepted: 18 May 2022
Published: 20 May 2022

1. Introduction

The rolling bearing plays an important role in rotating machinery, its health is related directly to the overall operating conditions and the quality of the mechanical equipment. Bearing failure can lead to equipment failure and cause serious economic losses or casualties to the enterprise. Therefore, it is very important to monitor and diagnose the health status of rolling bearings through their vibration data to ensure normal production in enterprises [1–7].

In recent years, more and more deep learning models have been developed and applied to the fault diagnosis of rolling bearings because of the end-to-end diagnosis ability. The typical CNN neural network is widely applied in the field of bearing-fault diagnosis because of its simplistic network structure and high accuracy diagnosis ability [8–11]. Although the feature extraction capabilities of the CNN model from the one-dimensional time-series signal or two-dimensional image can be increased by continuously stacking more convolutional layers, the CNN cannot capture long-range feature information [12,13]. To solve this problem, the position embedding method is introduced into the CNN to analyze the sequential nature of the time series signal or two-dimensional image, the multihead self-attention and parallel training mechanisms are all incorporated into the CNN model to capture the globally sensitive features from the time series signal or two-dimensional image quickly [14], thus the transformer model is proposed and applied in the field of natural language processing and image recognition [15]. After that, Yifei Ding

et al. [16] applied the transformer method to the field of fault diagnosis of mechanical equipment. However, the transformer has low computational efficiency and large memory consumption. In order to solve this problem, a vision transformer (ViT) which removes the decoder block of the transformer model is proposed for application in vision processing with a higher recognition performance, because it can not only inherit the multiheaded self-attention mechanism and relative position embedding method of the transformer but can also adopt the parallel learning mechanism and be prone to capture the global spatiotemporal information of an image [17–19]. Based on the advantages of ViT, a one-dimensional ViT architecture with multiscale convolution fusion is proposed to capture the fault features in multiple time scales with the transformer and achieve high diagnosis accuracy on the bearing fault dataset [20]. However, in the diagnosis process, ViT cannot thoroughly reveal the fault features concealed in the vibration signals, especially when fluctuations in the working conditions occur. This can affect the diagnosis performance of ViT. In addition, the diagnosis accuracy and generalization of the ViT model can be degraded because of over-fitting.

To capture more fault-related information, some time–frequency signal-processing methods, such as wavelet transform (WT), empirical mode decomposition (EMD), etc., can not only denoise the original vibration signal but also decompose the signal into different scale components which are combined with a deep learning model to extract fault features for improving the diagnosis accuracy [21,22]. However, the EMD and its variants suffer from mode mixing which decreases the decomposition performance. WT can overcome the problem, and discrete wavelet transform (DWT) can decompose the original vibration signal into the required scale components without reducing the amplitude, Continuous wavelet transform (CWT) can detect the singularity of the different scale components. Thus, the DWT over CWT can be utilized to detect the singularity of the required scale components for bearing fault diagnosis.

Integration learning has been widely applied to the fault diagnosis of bearings by flexibly fusing the preliminary diagnosis results of multiple base classifiers to obtain diagnosis results with higher accuracy and generalization ability because of the complementary classification behavior among different base classifiers. When integrated deep learning models combined with the different scale components of the original signal are utilized to diagnose the fault classes, higher diagnosis accuracy and generalization can be produced. Currently, some integrated deep learning models have been developed to apply to the field of fault diagnosis, and achieve good diagnosis results [23–25].

To the best of our knowledge, integrated learning has not been introduced into the ViT model to diagnose bearing faults. In order to improve the diagnosis accuracy and generalization of ViT, an integrated ViT model combined with wavelet transform and the soft voting method is proposed in this paper. The main contributions of the proposed diagnosis method are summarized as follows:

(1) The integrated ViT based on the soft voting fusion method is suggested to diagnose the bearing fault with high accuracy and generalization;
(2) DWT is used to decompose the original signal into different subsignals in different frequency bands and denoise the subsignals. After that, CWT is utilized to transform the subsignals into time–frequency representation (TFR) maps which can describe the singularity of the different subsignals;
(3) The ViT model can dig out more hidden fault-related information from the different TFR maps of the subsignals in different frequency bands.

The rest of the paper is organized as follows: Section 2 introduces the integrated ViT model which is combined with wavelet transform and the soft voting method; Section 3 presents the fault diagnosis flowchart of the integrated ViT; Section 4 gives the fault diagnosis experimental analysis of bearings based on the integrated ViT; lastly, the Conclusion is shown in Section 5.

2. Integrated Vision Transformer Model

The integrated ViT model not only inherits the advantages of integrated learning but also inherits the advantages of the ViT model, which can improve the diagnosis accuracy and generalization. Figure 1 shows the proposed fault diagnosis scheme diagram of the proposed integrated ViT model. Firstly, the vibration signal is gradually truncated through the sliding time window and divided into different data segmentations (data samples) which are decomposed into n subsignals in different frequency bands by DWT, and these subsignals are transformed into the corresponding n TFR maps by CWT, and then n individual ViT models which consist of a linear projection of flattened patches (embedding layer), the transformer encoder and MLP head are utilized to diagnose these n TFR maps of subsignals to obtain the preliminary diagnosis results, respectively. Finally, the final diagnosis decision can be obtained by the soft voting method used to fuse all the preliminary diagnosis results.

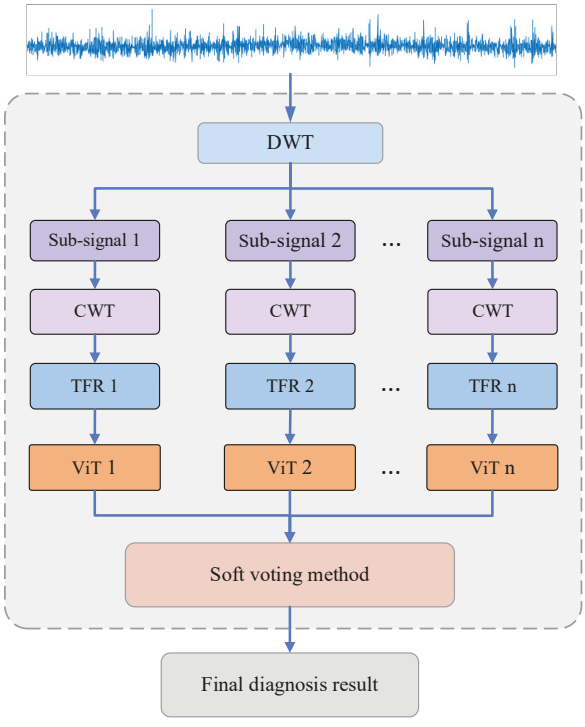


Figure 1. The diagnosis scheme diagram of the proposed integrated ViT.

2.1. DWT-Based Signal Decomposition

Referring to Figure 1, TFR maps of multiple subsignals in different frequency bands are input into multiple individual ViT models to diagnose the fault preliminarily. In order to reduce the influence of noise, DWT is used to decompose the original signal into different subsignals in different frequency bands without reducing the amplitude.

The discrete wavelet transform (DWT) can map any stationary or non-stationary signal to a set of base functions formed by wavelet scaling to obtain subsignals distributed in different frequency bands with complete information in the pass frequency range [26]. Based on the rules, the fault-related information in different frequency bands can be mined to diagnose the fault. Thus, through the scaling and translation of wavelet function basis

and scale function, the original signal can be decomposed into different subsignals with different scales. The detailed algorithm of DWT is described as follows.

(1) Given a time series signal a_0 of length N is decomposed by the Mallat tower wavelet decomposition algorithm [27], the decomposition process can be expressed as

$$\begin{cases} a'_{i+1} = Ha_i \\ d'_{i+1} = Ga_i \end{cases} \quad (1)$$

where H denotes the low-pass filter, G denotes the high-pass filter, a_i is the signal to be decomposed. a'_{i+1} and d'_{i+1} are the low-frequency and high-frequency coefficients obtained by the one-half down-sampling method, respectively.

(2) The obtained low-frequency coefficients can be decomposed repeatedly according to Equation (1). Thus, the low-frequency coefficients obtained from the decomposition of level j and the high-frequency coefficients obtained from the decomposition of each level are reconstructed to obtain subsignals in different frequency bands. For example, Figure 2 shows the 4-level decomposition result of the signal, A_4 and a set of subsignals D_4, D_3, D_2, D_1 represent approximate signal and detailed signals with frequency from low to high respectively. The relationship between the decomposed components (subsignals) and the original signal X can be expressed as

$$X = A_4 + D_4 + D_3 + D_2 + D_1 \quad (2)$$

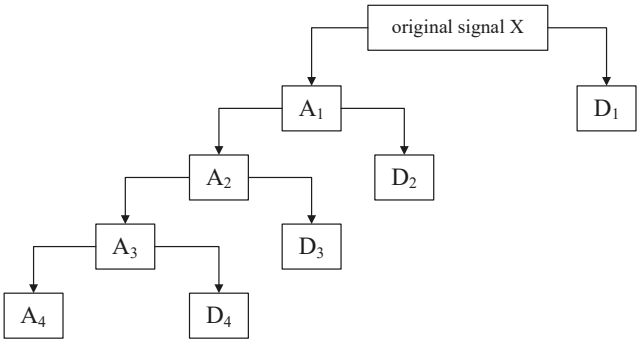


Figure 2. The DWT decomposition schematic. A_i indicates the ith layer approximate signal, D_i indicates the ith layer detailed signal.

2.2. Time–Frequency Analysis Based on CWT

The time–frequency analysis method is mainly used to reveal the time–frequency representation (TFR) of the subsignals in different frequency bands which can describe the relationship between the time and the frequency. At present, there are many time–frequency analysis methods used to analyze the time–frequency characteristics of vibration signals, such as short-time Fourier transform (STFT), Wigner–Ville distribution (WVD), and continuous wavelet transform (CWT). However, the STFT is unable to locate the time and frequency of non-stationary signals accurately [28]; the WVD is prone to frequency aliasing and cross-term interference [29]. In contrast, the CWT not only has good time–frequency resolution and time–frequency localization ability, but also detects the singularity of the signal. Thus, the corresponding time–frequency map of the subsignals in different frequency bands can depict the distinguished fault-related information [30,31]. The analysis process of CWT is described as follows.

Assume that the mother wavelet or basic wavelet function ψ is satisfied with $\psi \in L^1(R) \cap L^2(R)$ and $\hat{\psi}(0) \in 0$, the wavelet function family can be obtained by scaling and translation of function ψ. The wavelet function is written as follows

$$\psi_{a,b}(t) = |a|^{-1/2} \psi\left(\frac{t-b}{a}\right), a, b \in R, a \neq 0 \tag{3}$$

where $\{\psi_{a,b}\}$ is the analytic wavelet or continuous wavelet, a is the scaling factor of changing the wavelet shape, b is the translation factor of the wavelet shift. Thus, the CWT for an arbitrary function $f(t) \in L^2(R)$ can be expressed as

$$W_f(a,b) = \langle f, \psi_{a,b} \rangle = |a|^{-1/2} \int_R f(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt \tag{4}$$

where $\overline{\psi(t)}$ is the complex conjugate of $\psi(t)$, the symbol $\langle f, \psi_{a,b} \rangle$ is the inner product of the function f and ψ. W_f (a,b) denotes the coefficients of the wavelet function with scale a and offset b, which represent the similarity between the wavelet function and the original signal, and both a and b are continuous variables.

In order to obtain the fault-related TFR of a subsignal in different frequency bands in this paper, the CWT is used to transform the signal into a TFR, which is described as follows:

Assuming that f_s is the sampling frequency and F_c is the wavelet center frequency, the actual frequency F_a corresponding to scale a is written as

$$F_a = F_c \times f_s / a \tag{5}$$

In order to make the transformed frequency sequence an equal difference sequence, the scale sequence must take the following values.

$$c/totalscal, \cdots, c/(totalscal-1), c/4, c/2, c \tag{6}$$

where *totalscal* is the length of the scale series used in the wavelet transform of the signal, which is set as 256 here, and c is a constant.

On the basis of the sampling theorem, the actual frequency corresponding to the scale *c/totalscal* should be $f_s/2$. The value of the constant c can be calculated according to Equation (5), which can be obtained by the following equation

$$c = 2 \times F_c \times totalscal \tag{7}$$

Accordingly, the required scale sequence is obtained by substituting Equation (7) into Equation (6).

After determining the wavelet basis function and scale, the wavelet coefficients W_f (a,b) are obtained by applying the continuous wavelet transform principle of Equation (4). Then the scale sequence is converted into the actual frequency sequence f by Equation (5). Finally, the TFR map can be plotted.

2.3. Vision Transformer Model (ViT)

A transformer is a typical neural network model that relies entirely on a self-attention mechanism to establish the relationship between input and output, which can consider the global information comprehensively and be trained in parallel because of the parallel architecture that is completely different from the sequential structure of the traditional recurrent neural network (RNN). Figure 3 shows the architecture of the transformer model which mainly consists of a positional embedding layer, an encoder and a decoder. The positional embedding is used to add the relative positional information of the input data to the data processed by the embedding layer, thus, the transformer can better solve the long-time dependency problem. Based on these characteristics, the transformer can achieve

good performance on much vision detection, but it requires a good deal of memory and computational power.

In order to solve this problem, the vision transformer (ViT) was proposed by Dosovitskiy [17]. ViT has been applied widely to the field of image and vision recognition because of lower computational power and memory consumption, fewer training parameters and fewer training samples. Figure 4 shows the structure of the ViT model which consists of a linear projection of flattened patches (embedding layer), a transformer encoder and an MLP head. The model's first step is to divide an input image into a sequence of image patches. These image patches are then passed through a trained linear projection layer which plays the role of an embedding layer and outputs the vectors of fixed size. Position embeddings are linearly added to the sequence of image patches so that the images can retain their positional information. Then this new sequence of image patches is fed into the transformer encoder which is mainly composed of a multihead attention layer and a multilayer perceptron (MLP) layer; the multihead attention layer splits the inputs into several heads so that each head can learn different levels of self-attention. The outputs of all the heads are then concatenated and passed through the MLP head which is added to the transformer encoder to give the network's output classes.

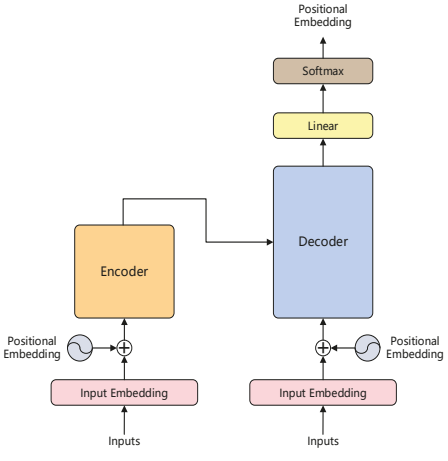


Figure 3. The architecture of transformer model.

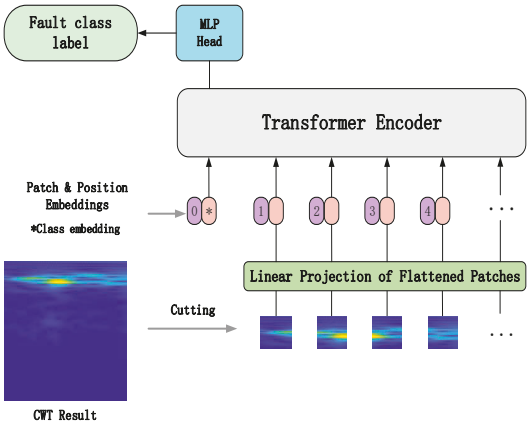


Figure 4. The structure of ViT model. * denotes the embedded class label vector.

2.3.1. Embedding Layer

The embedding layer is mainly for implementing the linear projection of flattened image patches and retaining the positional information and one-dimensional feature vector and class labels of the image patches. Suppose the input image $x \in R^{h \times w \times c}$, where h denotes the height of the image, w denotes the width of the image, and c denotes the number of channels of the image; the image is split into N image patches with length p and width p firstly, and then the image is flattened into a one-dimensional sequence $x_p \in R^{N \times (p*p*c)}$. After that, a linear projection is conducted on the one-dimensional sequence $x'_p \in R^{N \times D}$, these image patches are mapped into the D dimension vector space.

Additionally, the class label and the positional information of the image patch are all added to the outputs of embedding layer. Thus, a new sequence of image patches that contains the image features and the positional and class label information is obtained, which is the input of the transformer encoder.

2.3.2. Transformer Encoder

Each transformer encoder layer consists of multiple identical modular layers arranged in a stack, its internal structure is shown in Figure 5a. Each module layer contains two sublayers which are a multiheaded self-attention layer and MLP feed-forward network respectively, the structure of the MLP feed-forward network can be seen in Figure 5b. To improve the accuracy of the network model by increasing the depth of the network generally, each sublayer is internally connected using residuals, and layer normalization is used at the end of each sublayer to improve the training speed and generalization performance of the neural network. The output of each sublayer can be expressed as

$$o = LayerNorm(x + Sublayer(x)) \tag{8}$$

where $Sublayer(x)$ indicates the multiheaded self-attentive function and MLP function in each sublayer, $LayerNorm(\blacksquare)$ denotes the normalization function.

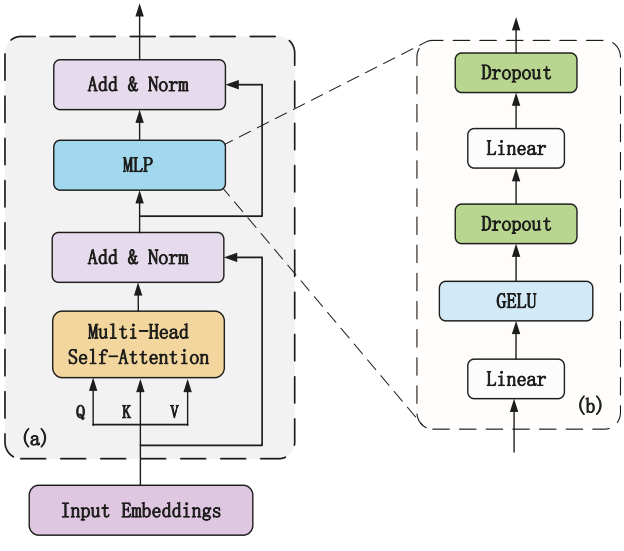


Figure 5. The architecture of the transformer encoder module. (**a**) the architecture of the transformer encoder module, (**b**) the internal architecture of the MLP.

The internal structure of the MLP layer and the multihead self-attention layer are described as follows.

- MLP layer

The internal structure of the MLP is shown in Figure 5b, which comprises a fully connected layer, GELU function, and dropout function. To improve the convergence of the network, in the feedforward layer the ViT uses the Gaussian error linear unit (GELU) activation function instead of the ReLU activation used in transformer. The output of the GELU activation is expressed as follows

$$GeLU(x) = x \cdot \frac{1}{2}\left[1 + erf\left(x/\sqrt{2}\right)\right] \tag{9}$$

where x denotes the input, and $erf(\blacksquare)$ denotes the Gaussian error function.

- Multiheaded self-attention layer

The self-attention mechanism allows the network model to extract local valid features, but a single attention mechanism can only learn relevant information in one representation space. To synthetically extract long-distance features from a global image, a multiheaded self-attention mechanism is used to jointly focus on features from different representation subspaces at different locations. The structure of the multihead attention layer is shown in Figure 6. The self-attention mechanism uses scaled dot-product attention to calculate the attention value of the feature matrix. The calculation formula of the scaled dot-product attention is written as follows [32],

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{10}$$

$$Q = X_f W^Q \tag{11}$$

$$K = X_f W^K \tag{12}$$

$$V = X_f W^V \tag{13}$$

where Q is the query matrix, K is the key matrix, V is the value matrix. These three matrices are obtained by multiplying the input feature matrix X_f with the parameter matrices W^Q, W^K and W^V respectively, d is the dimension of Q, K and V.

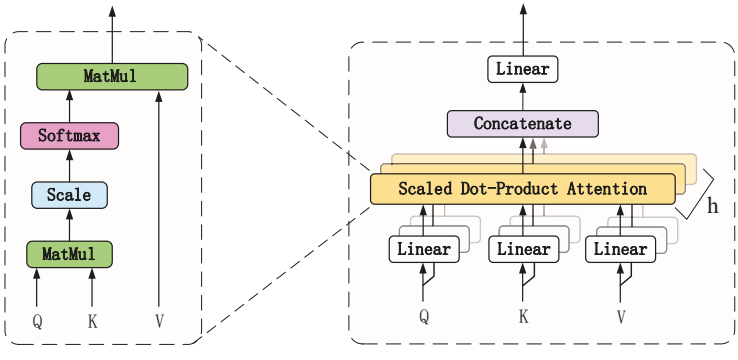


Figure 6. The architecture of multilayer attention layer.

The multihead self-attention mechanism is a combination of multiple self-attention mechanisms, which use multiple self-attention heads to learn features from different representation subspaces, respectively, and then the multiple attention value is combined and transformed linearly, thus the final attention value is obtained to realize the representa-

tion under these different constraint conditions. The multihead self-attention mechanism equation can be expressed as follows

$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_h) \cdot W^O \tag{14}$$

$$head_i = Attention\left(Q \cdot W_i^Q \cdot K \cdot W_i^K, V \cdot W_i^V\right) \tag{15}$$

where W_i^Q, W_i^K, W_i^V are the weight matrices of the ith attention head Q, K and V, W^O is the weight matrix of multihead attention, h is the number of attention heads, *Concat* function is to concatenate the output values of each attention head.

2.3.3. MLP Head

Generally, the standard MLP head layer of ViT consists of a fully connected layer and activation function, which is used to diagnose the fault classes. In order to reduce the calculation workload of the ViT model, a Gaussian error linear unit (GELU) activation function is adopted in this paper. Thus, the input data processed by the transformer encoder layer is input into the MLP head to obtain the probability value of each fault class, and the final fault class can be obtained according to the maximum probability value.

2.4. Decision Fusion Based on Soft Voting Method

When the output of a single classifier is the probability value of each fault class, the fusion is most generally performed by the soft voting method [33]. Considering the classification output of ViT is the probability value corresponding to each fault class, the soft voting method is adopted to fuse all the outputs of multiple ViTs to obtain the final diagnosis results. The fusion process based on the soft voting method is shown in Figure 7. Suppose the output probability vectors $y_k(x^k)$ of the time–frequency map x^k produced by the kth base classifier $\{M^{(k)}\}$, the maximum value $Y(X)$ is taken as the final classification result, which is defined as follows

$$Y(X) = max\left\{\frac{1}{K}\sum_{k-1}^{K} y_k\left(x^k\right)\right\} \tag{16}$$

where x^k denotes the CWT time–frequency map of the subsignal in the kth frequency bands decomposed by DWT on the original data samples x, $max()$ denotes the maximum function, K is the number of base classifiers.

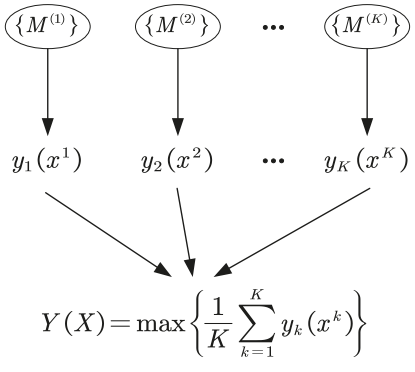


Figure 7. The fusion process of the soft voting method.

3. Diagnosis Method Based on Integrated ViT Model

The diagnosis flowchart based on the proposed integrated ViT model is shown in Figure 8. The collected vibration signal is segmented into different data samples by the

sliding time window, and then these data samples are divided into a training dataset and test dataset. By the DWT and CWT method, the data samples in the training dataset are decomposed into different subsignals in different frequency bands to obtain the different time–frequency representation (TFR) maps which are input into individual ViT models respectively, and then the multiple trained ViT models can be obtained. In the same way, the TFR maps of different subsignals in different frequency bands in the test dataset are also obtained, which are input to the multiple trained ViT models, respectively, to obtain the preliminary diagnosis results. After that, the final diagnosis result is obtained by using the soft voting method to fuse all the preliminary diagnosis results.

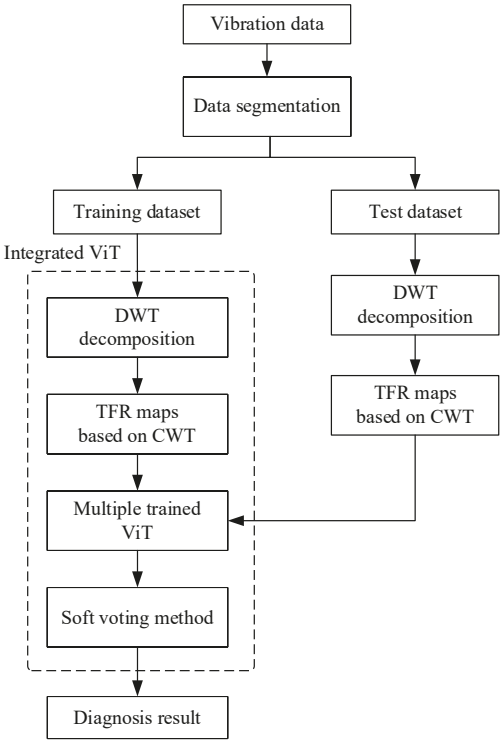


Figure 8. The diagnosis flowchart based on integrated ViT.

In addition, the number of the individual ViT model is determined by the number of the subsignals in different frequency bands, which is set as five here. The loss functions of all ViT models are all cross entropy loss functions which are written as follows:

$$Loss = -\frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} \cdot log(p_{i,k}) \tag{17}$$

where K is the number of fault classes, N is the number of training samples, $y_{i,k}$ is the symbolic function (0 or 1), which takes 1 if the true fault class of sample i is equal to k, take 1, otherwise it is 0. $p_{i,k}$ denotes the probability value of fault class k that the data sample i belongs to. The parameters of each individual ViT model are trained by the TFR maps, respectively.

4. Fault Diagnosis of Rolling Bearing

4.1. Acquisition of Bearing Vibration Signal

In order to verify the effectiveness of the proposed integrated ViT model for the fault diagnosis of the rolling bearing, the diagnosis method was utilized to diagnose the fault signals obtained from the Case Western Reserve University (CWRU) Bearing Data Center [34,35]. As shown in Figure 9, the experimental equipment comprised a motor, rolling bearing, torque sensor, and dynamometer. The bearing under test was 6205-2RS JEM KSF, a deep groove ball bearing. The drive end (DE) bearing fault signal with a sampling frequency 48 kHz, a spindle speed 1797 r/min and a load 0 hp was collected.

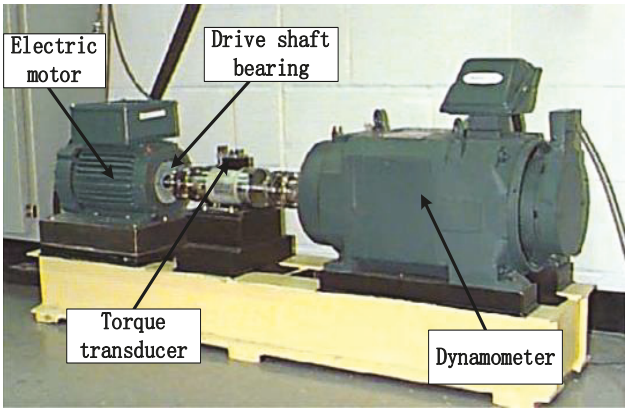


Figure 9. Experimental setup of rolling bearing fault.

The electrical discharge machining (EDM) method was used to simulate the different fault categories and severity of the bearings. Three different diameters of single point damage (0.18 mm, 0.36 mm and 0.53 mm) on the inner ring, outer ring and rolling element of the bearing were introduced, respectively. The statistics of the dataset are described in Table 1. The dataset contains 10 fault classes, the number of data samples for each fault class was 500, of which 350 were training samples and 150 were test samples. Thus, the total number of training data samples and test samples were 3500 and 1500, respectively. In addition, the number of data points per data sample was 1024.

Table 1. The statistics of bearing fault dataset.

Fault Class Conditions	Class Label	The Number of Training Samples	The Number of Test Samples	Fault Size (mm)
Normal	1	350	150	0
Slight inner ring	2	350	150	0.18
Medium inner ring	3	350	150	0.36
Severe inner ring	4	350	150	0. 53
Slight outer ring	5	350	150	0.18
Medium outer ring	6	350	150	0.36
Severe outer ring	7	350	150	0. 53
Slight rolling element	8	350	150	0.18
Medium rolling element	9	350	150	0.36
Severe rolling element	10	350	150	0.53

4.2. Wavelet Transform Analysis of Vibration Signal

4.2.1. Obtaining Subsignals in Different Frequency Bands Based on DWT

Considering that the shape of Daubechies (Db) wavelet function is similar to the waveform of bearing vibration signal, and in order to obtain a better frequency band division effect and reduce the calculation time, DWT based on Db5 wavelet function was used to decompose the data sample into different subsignals in frequency bands in this paper. Figure 10 shows the four level decomposition results of the bearing vibration signal which are the original signal, detail signals D1, D2, D3, D4 and approximate signal A4, respectively. It can be seen that the different detail subsignals and approximate subsignals depict the vibration characteristics from different scales such as the vibration amplitudes and frequency, but the relationship between time and frequency cannot be shown.

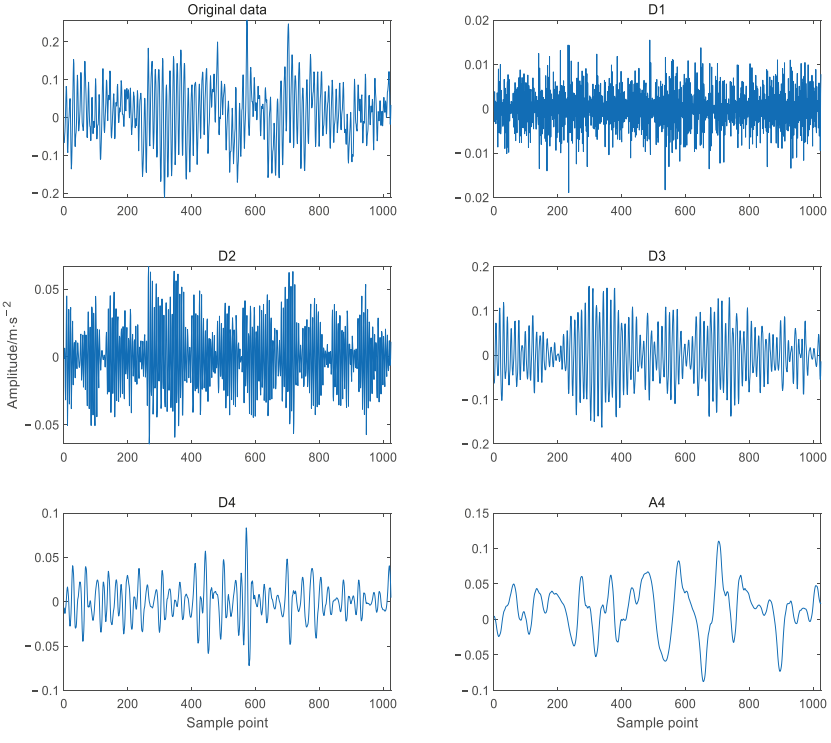


Figure 10. Decomposition of vibration signal based on DWT.

4.2.2. Time–Frequency Analysis Based on CWT

In order to obtain the TFR maps of subsignals in different frequency bands to describe the relationship between time and frequency, the different detail subsignals D1, D2, D3, D4 and the approximation subsignal A4 was transformed by the CWT based on the cmor3-3 wavelet basis function which was selected because its shape is similar to the impact signal of a bearing fault. Figure 11 shows the TFR maps of the original vibration signal and different approximate and detail subsignals. From the figure, it can be seen that the frequency components of the original vibration signal contained the frequency components of approximate subsignals A4 and detail subsignals D1 and D2, but the frequency components of detail subsignals D1 and D2 were obviously different from the frequency components of the original vibration signal, namely that these detail and approximate subsignals could reveal more frequency information on the bearing, and they could depict the relationship

between time and frequency. All these demonstrated that the CWT method can dig out more fault-related information from the original vibration signal.

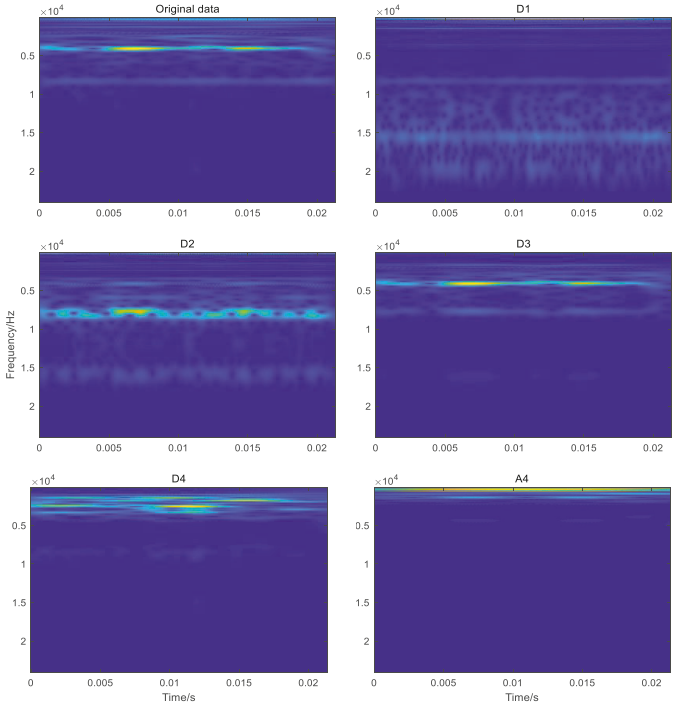


Figure 11. TFR maps of subsignal and original vibration signal.

4.3. Diagnosis Analysis

In order to verify the effectiveness of the integrated ViT models based on wavelet transform and the soft voting method, the TFRs of five different detail and approximation subsignals in different frequency bands of size $64 \times 64 \times 3$, were input to five individual ViT models, respectively, to preliminarily diagnose the fault of the bearing, and then the soft voting method was used to fuse all the preliminary diagnosis results to obtain the final diagnosis result.

Figure 12 shows the diagnosis results using the individual ViT model with different detail and approximation subsignals, respectively, decomposed from some vibration data samples, where the *X*-axis denotes the test data sample number and *Y*-axis denotes the fault-class labels. From the figure, it can be seen that the diagnosis accuracy produced by ViT model with the detail subsignal D1 achieved 95.07%, which is the highest among all diagnosis accuracy produced by all the individual ViT models with other detailed and approximate subsignals; the diagnosis accuracy produced by the individual ViT model with the detailed subsignal D2, D3 and D4 was 94.80%, 76.47%, respectively, the diagnosis accuracy produced by the individual ViT model with approximate subsignal A4 was 74.40%, which was only higher than that of the individual ViT model with D4, the diagnosis accuracy of the individual ViT model with D4 was the lowest, only 73.73%. This is mainly because the different detailed and approximate subsignals in the different frequency bands contained different amounts of fault-related information, and the amount of fault information in different frequency bands can affect the diagnosis accuracy of the individual ViT model directly.

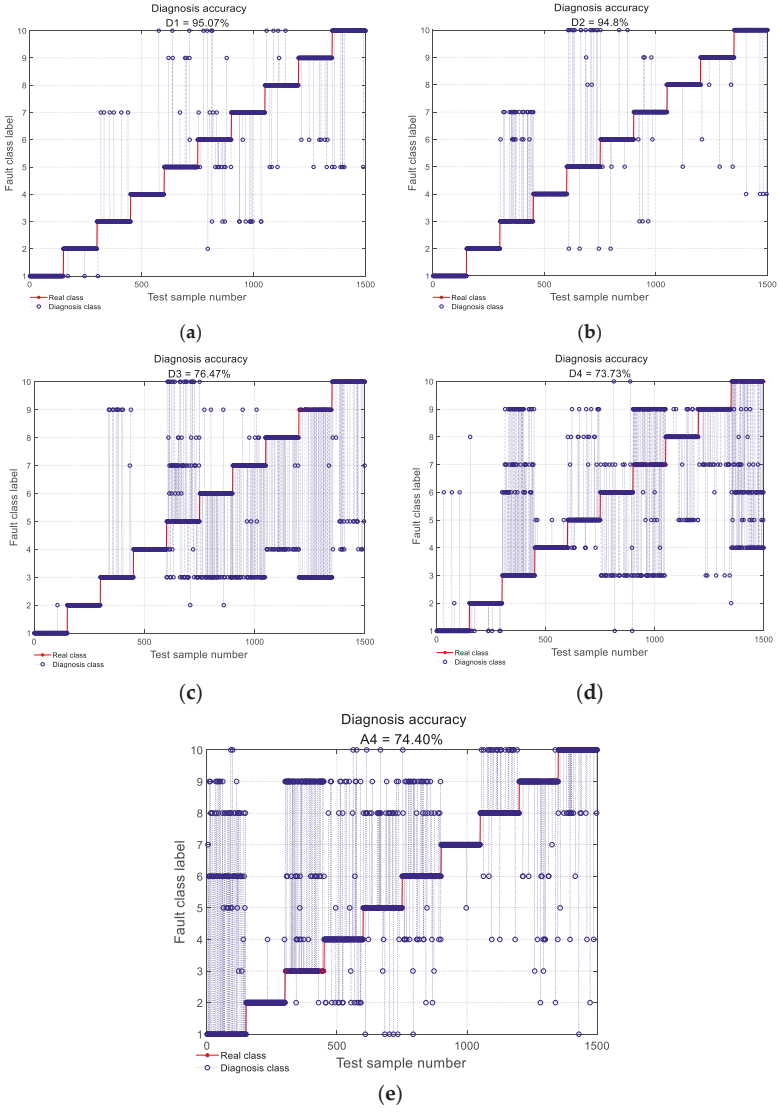


Figure 12. Diagnosis results of an individual ViT with different detail and approximation subsignal TFR. (**a**–**e**) Diagnostic accuracy with an individual ViT with D1, D2, D3, D4 and A4 subsignal TFR, respectively.

The final diagnosis results of the integrated ViT model using the soft voting method for decision making to fuse all the preliminary results of five individual ViT models with five different detailed and approximate subsignals in different frequency bands were obtained and are shown in Figure 13. In addition, to further validate the effect of integrated ViT, the diagnosis results of the integrated ViTs with the different numbers of individual ViT models are also shown in Figure 13, where the *X*-axis and the *Y*-axis indicate the names of the detailed and approximate subsignals and the diagnostic accuracy, respectively; the histogram indicates the diagnosis accuracy of the individual ViT with different TFR maps of the subsignals in different frequency bands, the curve indicates the diagnostic accuracy

of the integrated ViT model with the first n TFR maps of subsignals in different frequency bands. From the figure, it can be seen that the diagnosis accuracy of the integrated ViT models increased with the number of individual ViT models involved in the integration. The diagnosis accuracy of the integrated ViT models with first two individual ViT models can reach 99.13%, and the diagnosis accuracy of integrated ViT with the five individual ViT models achieved 100.00%, which exceeded the accuracy of all individual ViT models and the accuracy of other integrated ViT models with different numbers of individual ViT models. The diagnostic accuracy of all the integrated ViT models with different numbers of individual ViT models was consistently higher than the highest diagnostic accuracy of the individual ViT model. All these indicate that the integrated ViT model based on the soft voting method has a superior diagnosis performance to the individual ViT model, integrated learning can improve the diagnosis accuracy of the individual ViT model.

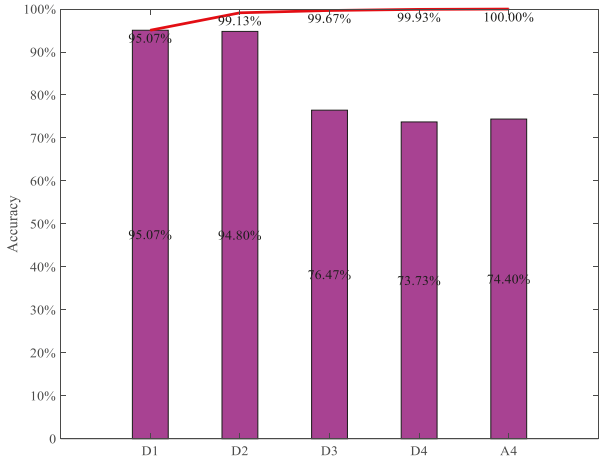


Figure 13. Diagnosis accuracy of the integrated ViT models with first n ViT. The histogram indicates the diagnosis accuracy of the individual ViT with different TFR maps respectively, the red curve indicates the diagnostic accuracy of the integrated ViT model with the first n TFR maps of subsignals.

4.3.1. Comparison with Other Integrated Models and Individual Models

To verify the superiority of the integrated ViT model, the TFRs of the subsignal in different frequency bands were also input into the integrated CNN models based on the soft voting method to diagnose the bearing fault. The structure parameters of the individual CNN are shown in Table 2.

Table 2. The parameters of CNN model.

Layer	Input Size	Output Size
Conv2D	64, 64, 3	64, 64, 32
Conv2D	64, 64, 32	64, 64, 32
MaxPooling2D	64, 64, 32	32, 32, 32
Flatten	32, 32, 32	32,768
Dense	32,768	32
Dense	32	10

Figure 14 shows the diagnosis accuracy of the integrated CNN models using the different numbers of individual CNN models with the TFRs of subsignals in different frequency bands; the histogram indicates the diagnosis accuracy of the individual CNN

with different TFR maps of subsignals in different frequency bands, the curve indicates the diagnosis accuracy of the integrated CNN models using the first n individual CNN models with the TFR of different subsignals. It can be seen that the diagnosis accuracy of the integrated CNN models increased with the number of individual CNN models involved in the integration. The highest diagnosis accuracy obtained by the integrated CNN model using the different numbers of individual CNN models was 99.13%, which was higher than the highest diagnosis accuracy of the individual CNN model, i.e., 96.20%. From Figures 12 and 13, it can be seen that the diagnosis accuracy of the integrated CNN model was always lower than that of the integrated ViT model with the same number of individual diagnosis models, and the highest diagnosis accuracy of integrated CNN model was lower than the highest diagnosis accuracy of the integrated ViT model. This could indicate that the integrated ViT has superior diagnosis ability.

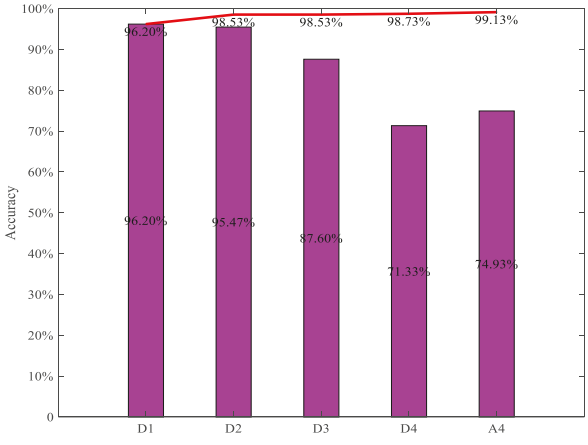


Figure 14. Diagnosis accuracy of integrated CNN using the first n individual CNN. The histogram indicates the diagnosis accuracy of the individual ViT with different TFR maps respectively, the red curve indicates the diagnostic accuracy of the integrated ViT model with the first n TFR maps of subsignals.

In addition, to verify the stability of the proposed integrated ViT model, the five diagnosis tests were conducted by the individual ViT model, the integrated ViT model and the integrated CNN model, respectively. Table 3 shows the average diagnostic accuracy, the minimum and maximum accuracy produced by the three diagnosis models. From the table, it can be seen that the mean, minimum and maximum values of the diagnosis accuracy of the integrated ViT model were 99.87, 99.47 and 100.00%, respectively, which were all the highest among the corresponding accuracies of three diagnosis models, and the mean, minimum and maximum values of diagnosis accuracy of the integrated CNN model were higher than those of the individual ViT model, respectively. All these demonstrate that the integrated ViT model has higher diagnosis accuracy and diagnosis stability compared with the integrated CNN and individual ViT, and integrated learning can further improve the diagnosis accuracy and stability of the individual ViT.

Table 3. Performance comparison of individual ViT, integrated ViT and integrated CNN.

Diagnostic Model	Mean of Diagnosis Accuracy	Minimum of Diagnosis Accuracy	Maximum of Diagnosis Accuracy
ViT	98.73%	97.76%	99.87%
Integrated ViT model	99.87%	99.47%	100.00%
Integrated CNN model	99.13%	98.53%	99.87%

4.3.2. Generalization Analysis of the Integrated ViT

To validate the generalization of the proposed integrated ViT model based on the soft voting method, the fault diagnosis analysis of the three diagnosis models was conducted on three different datasets under three different working conditions (0 hp load and 1797 rpm, 1 HP load and 1772 rpm and 2 HP load and 1750 rpm) which are referred to as dataset 1, dataset 2 and dataset 3, for convenience. Each dataset contained 10 fault classes, the number of samples for each fault class was 200, including 140 training samples and 60 test samples. Thus, each dataset had 1400 training samples and 600 test samples. Each sample had 1024 data points. Table 4 shows the diagnosis results of the integrated ViT, integrated CNN and individual ViT on the three different datasets.

Table 4. Diagnosis results of three diagnosis models on three datasets.

Diagnosis Model	Diagnosis Accuracy (%)		
	Dataset 1	Dataset 2	Dataset 3
Integrated ViT	100.00	99.67	99.83
Integrated CNN	99.17	99.33	98.33
ViT	98.83	98.67	97.83

From Table 4, it can be seen that the diagnosis accuracy of the integrated ViT on the three datasets was 100.00, 99.67 and 99.83%, respectively; the maximum difference among these three diagnosis accuracies is 0.33%, and the minimum difference among them is 0.17%. The diagnosis accuracy of the integrated CNN on the three datasets was 99.17, 99.33 and 98.33%, respectively; the maximum difference among these three average diagnosis accuracies is 1%, and the minimum difference among them is 0.16%. The diagnosis accuracy of the individual ViT on the three datasets was 98.83, 98.67 and 97.83%, the maximum difference among these three diagnoses accuracies is 1%, and the minimum difference among them is 0.16%. The diagnosis accuracies of the integrated ViT on the three datasets are the highest among the three diagnosis models respectively, the diagnosis accuracies of the individual ViT on the three datasets are the lowest among the three diagnosis models respectively. In addition, the maximum difference of the diagnosis accuracy of the integrated ViT on the three datasets was the lowest among the three diagnosis models, the minimum difference of the diagnosis accuracy of the integrated ViT was only 0.01% higher than that of integrated CNN and individual ViT, respectively. All these can demonstrate that the integrated ViT has stronger diagnosis generalization than the integrated CNN and individual ViT, and furthermore, has the highest diagnosis accuracy among the three methods.

5. Conclusions

This paper proposes an integrated ViT model with the TFR maps of subsignals in different frequency bands based on the soft voting method to diagnose bearings. In the diagnosis process, DWT is used to decompose the vibration signal into different subsignals in different frequency bands and CWT is utilized to obtain TFRs of subsignals in different frequency bands, and then the TFR maps of the different subsignals are input into multiple individual ViT models to diagnose the fault preliminarily, and lastly, the final diagnosis result is obtained by the fusion method based on the soft voting method.

The effectiveness and generalization of the proposed integrated ViT model were verified by comparison with the integrated CNN model based on the soft voting method and individual ViT model. Through a multifaceted comparison of the three methods on different experimental datasets, the diagnosis results demonstrated that the proposed integrated ViT has higher diagnosis accuracy and generalization than the integrated CNN and individual CNN model for fault diagnosis of rolling bearings. All these show that the integrated ViT has a promising development prospect in the field of fault diagnosis of mechanical equipment. However, it was found that the number of ViT models used for integrated learning affected the diagnosis accuracy in the process of fault diagnosis, so how to select the number of ViT models with subsignals in different frequency bands will be studied further in future.

Author Contributions: Conceptualization, Z.X.; methodology, X.T. and Z.X.; software, X.T.; validation, Z.X. and Z.W.; formal analysis, X.T.; investigation, X.T. and Z.X.; data curation, X.T.; writing—original draft preparation, X.T.; writing—review and editing, Z.X.; visualization, X.T.; supervision, Z.W.; project administration, Z.X.; funding acquisition, Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 51775391, the Open Research Foundation of State Key Lab. of Digital Manufacturing Equipment & Technology in Huazhong University of Science & Technology under Grant DMETK F2017010.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support this study are available at the website http://csegroups.case.edu/bearingdatacenter/pages/download-data-file (accessed on 24 August 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Glowacz, A. Ventilation diagnosis of angle grinder using thermal imaging. *Sensors* **2021**, *21*, 2853. [CrossRef] [PubMed]
2. Glowacz, A.; Tadeusiewicz, R.; Legutko, S.; Caesarendra, W.; Irfan, M.; Liu, H.; Brumercik, F.; Gutten, M.; Sulowicz, M.; Daviu, J.A.; et al. Fault diagnosis of angle grinders and electric impact drills using acoustic signals. *Appl. Acoust.* **2021**, *179*, 108070. [CrossRef]
3. Zhang, X.; Zhang, M.; Xiang, Z.; Mo, J. Research on diagnosis algorithm of mechanical equipment brake friction fault based on MCNN-SVM. *Measurement* **2021**, *186*, 110065. [CrossRef]
4. Pandya, D.H.; Upadhyay, S.H.; Harsha, S.P. Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. *Expert Syst. Appl.* **2013**, *40*, 4137–4145. [CrossRef]
5. Liguori, A.; Armentani, E.; Bertocco, A.; Formato, A.; Pellegrino, A.; Villecco, F. Noise reduction in spur gear systems. *Entropy* **2020**, *22*, 1306. [CrossRef]
6. Wang, Y.; Li, S.; Jia, F.; Shen, J. Multi-Domain Weighted Transfer Adversarial Network for the Cross-Domain Intelligent Fault Diagnosis of Bearings. *Machines* **2022**, *10*, 326. [CrossRef]
7. Ahmed, H.O.; Nandi, A.K. Intrinsic Dimension Estimation-Based Feature Selection and Multinomial Logistic Regression for Classification of Bearing Faults Using Compressively Sampled Vibration Signals. *Entropy* **2022**, *24*, 511. [CrossRef]
8. Pan, H.; He, X.; Tang, S.; Meng, F. An Improved Bearing Fault Diagnosis Method using One-Dimensional CNN and LSTM. *Stroj. Vestn. J. Mech. Eng.* **2018**, *64*, 443–452.
9. Wang, X.; Mao, D.; Li, X. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement* **2021**, *173*, 108518. [CrossRef]
10. Hasan, M.J.; Sohaib, M.; Kim, J.M. 1D CNN-based transfer learning model for bearing fault diagnosis under variable working conditions. In Proceedings of the International Conference on Computational Intelligence in Information System, Gadong, Brunei Darussalam, 16–18 November 2018; Springer: Cham, Switzerland, 2018; pp. 13–23.
11. Zhao, B.; Zhang, X.; Li, H.; Yang, Z. Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions. *Knowl. Based Syst.* **2020**, *199*, 105971. [CrossRef]
12. Zhang, S.; Zhang, S.; Wang, B.; Habetler, T.G. Deep learning algorithms for bearing fault diagnostics—A comprehensive review. *IEEE Access* **2020**, *8*, 29857–29881. [CrossRef]
13. Mon, Y.J. Vision Robot Path Control Based on Artificial Intelligence Image Classification and Sustainable Ultrasonic Signal Transformation Technology. *Sustainability* **2022**, *14*, 5335. [CrossRef]

14. Wang, H.; Xu, J.; Yan, R.; Sun, C.; Chen, X. Intelligent bearing fault diagnosis using multi-head attention-based CNN. *Procedia Manuf.* **2020**, *49*, 112–118. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
16. Ding, Y.; Jia, M.; Miao, Q.; Cao, Y. A novel time–frequency Transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mech. Syst. Signal Process.* **2022**, *168*, 108616. [CrossRef]
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26 April–1 May 2020.
18. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F.E.H.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
19. Chen, C.-F.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Addis Ababa, Ethiopia, 1–7 May 2021; pp. 357–366.
20. Weng, C.; Lu, B.; Yao, J. A One-Dimensional Vision Transformer with Multiscale Convolution Fusion for Bearing Fault Diagnosis. In Proceedings of the 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing), Nanjing, China, 15–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
21. Zhao, M.; Kang, M.; Tang, B.; Pecht, M. Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4290–4300. [CrossRef]
22. Chen, K.; Zhou, X.C.; Fang, J.Q.; Zheng, P.F.; Wang, J. Fault feature extraction and diagnosis of gearbox based on EEMD and deep briefs network. *Int. J. Rotating Mach.* **2017**, *5*, 1–10. [CrossRef]
23. Li, X.; Jiang, H.; Niu, M.; Wang, R. An enhanced selective ensemble deep learning method for rolling bearing fault diagnosis with beetle antennae search algorithm. *Mech. Syst. Signal Process.* **2020**, *142*, 106752. [CrossRef]
24. Zhang, Y.; Wang, J.; Zhang, F.; Lv, S.; Zhang, L.; Jiang, M.; Sui, Q. Intelligent fault diagnosis of rolling bearing using the ensemble self-taught learning convolutional auto-encoders. *IET Sci. Meas. Technol.* **2022**, *16*, 130–147. [CrossRef]
25. Xu, G.; Liu, M.; Jiang, Z.; Söffker, D.; Shen, W. Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. *Sensors* **2019**, *19*, 1088. [CrossRef]
26. Bruce, L.M.; Koger, C.H.; Li, J. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2331–2338. [CrossRef]
27. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
28. Feng, Z.; Liang, M.; Chu, F. Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples. *Mech. Syst. Signal Process.* **2013**, *38*, 165–205. [CrossRef]
29. Manap, M.; Abdullah, A.R.; Nikolovski, S.; Sutikno, T.; Jopri, M.H. An improved smooth-windowed wigner-ville distribution analysis for voltage variation signal. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 4982. [CrossRef]
30. Timoshevskaya, O.; Londikov, V.; Andreev, D.; Samsonenkov, V.; Klets, T. Digital Data Processing Based on Wavelet Transforms. In *Environment, Technologies, Resources, Proceedings of the 13th International Scientific and Practical Conference, Rezekne, Latvia, 17–18 June 2021*; Rezekne Academy of Technologies: Rēzekne, Latvia, 2021; Volume 2, pp. 174–180.
31. Li, P.; Yuan, H.; Wang, Y.; Chen, X. Pumping unit fault analysis method based on wavelet transform time-frequency diagram and cnn. *Int. Core J. Eng.* **2020**, *6*, 182–188.
32. Yan, G.; Liang, S.; Zhang, Y.; Liu, F. Fusing transformer model with temporal features for ECG heartbeat classification. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; IEEE: Piscataway, NJ, USA; pp. 898–905.
33. Rojarath, A.; Songpan, W.; Pong-inwong, C. Improved ensemble learning for classification techniques based on majority voting. In Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 26–28 August 2016; pp. 107–110.
34. The Case Western Reserve University Bearing Data Center. Bearing Data Center Fault Test Data. 1998. Available online: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file (accessed on 24 August 2021).
35. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64*, 100–131. [CrossRef]

sensors

MDPI

Article

Performance Degradation Prediction Using LSTM with Optimized Parameters

Yawei Hu [1], Ran Wei [2], Yang Yang [3,*], Xuanlin Li [1], Zhifu Huang [1], Yongbin Liu [1,*], Changbo He [1] and Huitian Lu [4]

[1] College of Electrical Engineering and Automation, Anhui University, Hefei 230601, China; yhu@ahu.edu.cn (Y.H.); z21301130@stu.ahu.edu.cn (X.L.); hzf125521@163.com (Z.H.); changbh@ahu.edu.cn (C.H.)
[2] Anhui NARI Jiyuan Electric Co., Ltd., Hefei 230601, China; weiran_nari@163.com
[3] China North Vehicle Research Institute, Beijing 100071, China;
[4] JJL College of Engineering, South Dakota State University, Brookings, SD 57007, USA; huitian.lu@sdstate.edu
* Correspondence: yangyang913@163.com (Y.Y.); lyb@ustc.edu.cn (Y.L.)

Abstract: Predicting the degradation of mechanical components, such as rolling bearings is critical to the proper monitoring of the condition of mechanical equipment. A new method, based on a long short-term memory network (LSTM) algorithm, has been developed to improve the accuracy of degradation prediction. The model parameters are optimized via improved particle swarm optimization (IPSO). Regarding how this applies to the rolling bearings, firstly, multi-dimension feature parameters are extracted from the bearing's vibration signals and fused into responsive features by using the kernel joint approximate diagonalization of eigen-matrices (KJADE) method. Then, the between-class and within-class scatter (SS) are calculated to develop performance degradation indicators. Since network model parameters influence the predictive accuracy of the LSTM model, an IPSO algorithm is used to obtain the optimal prediction model via the LSTM model parameters' optimization. Finally, the LSTM model, with said optimal parameters, was used to predict the degradation trend of the bearing's performance. The experiment's results show that the proposed method can effectively identify the trends of degradation and performance. Moreover, the predictive accuracy of this proposed method is greater than that of the extreme learning machine (ELM) and support vector regression (SVR), which are the algorithms conventionally used in degradation modeling.

Keywords: performance degradation; degradation prediction; KJADE; LSTM; IPSO; rolling bearing

Citation: Hu, Y.; Wei, R.; Yang, Y.; Li, X.; Huang, Z.; Liu, Y.; He, C.; Lu, H. Performance Degradation Prediction Using LSTM with Optimized Parameters. *Sensors* **2022**, *22*, 2407. https://doi.org/10.3390/s22062407

Academic Editor: Roberto Teti

Received: 5 February 2022
Accepted: 10 March 2022
Published: 21 March 2022

1. Introduction

Studies have shown that more than 45% of equipment failures in rotating machinery are caused by bearing failure [1]. The financial losses and major safety accidents that this causes in the industry can be avoided by assessing the deterioration status of equipment, which would bolster an organization's ability to detect faults in machine bearings [2].

According to previous studies on the topic, data-driven modeling has, over time, gradually become the most effective forecasting method [3,4]. In order to predict the remaining useful life (RUL) of bearings, a large number of studies, focusing on data-driven modeling, have been carried out, including the support vector machine (SVM) and artificial neural network (ANN) [5,6]. Zheng et al. proposed the ensemble SVM for the fault detection and diagnosis of rolling bearings, in which composite multiscale fuzzy entropy was used to discern health indicators [7]. However, traditional machine learning methods like SVM require a priori knowledge of feature engineering, which is extremely difficult to implement with regard to bearings due to the complex working conditions they operate under. Deep learning-based algorithms provide an alternative solution to this problem [8–10]. Chen et al. proposed a method based on neuro-fuzzy systems (NFSs) and Bayesian algorithms, which use trained NFSs as predictors to discern the degradation of a given machine's fault state over time [11]. Ren et al. proposed the use of the spectral

principal energy vector method in obtaining bearing signal feature vectors. They adopt the deep convolutional neural network to predict the RUL of rolling bearings [12]. The recurrent neural network (RNN), as an important subfield of deep learning, performs well with regarding time series processing because the RNN can forecast using all available historical data [13]. Malhi et al. made further strides towards putting forth a long-term prediction model for machine condition monitoring based on RNN [14]. However, the disappearance, or even the explosion, of gradients during network training seems to restrict this method's applicability [15].

To solve the issue, Hochreiter and Schmidhuber presented a variant of the RNN network in 1997, namely, the long short-term memory (LSTM) neural network, which addressed the problem by adding a gating mechanism [16]. At present, LSTMs are widely used in a variety of different fields, such as speech recognition, time series modeling, video classification, traffic flow prediction, and so on. Besides this, the LSTM method has also been used to predict bearing degradation, exploring the correlation between bearing degradation data and time [17]. Liu et al. proposed the use of an end-to-end model in predicting the degradation trends of bearings. His model used CNN for data reduction and feature extraction and a LSTM for time series processing [18]. Elsheikh et al. proposed bidirectional rocking long-term and short-term memory to predict the RUL of turbofan engines [19]. Tang et al. used a stacked automatic encoder (SAE) to obtain the bottleneck characteristics of bearing signals and predicted bearing performance degradation with a LSTM [20].

In application, the choice of the network structure, the number of hidden layers, and the learning rate setting will significantly influence the predictive capability of LSTMs [21]. Typically, the complex structure and parameters of LSTM neurons are mostly determined by experience, or by multiple parameter adjustments with expensive time, which involves a lot of randomness and reduces the model's predictive capability. Therefore, a set of hyper-parametric optimization algorithms were developed to select the parameters automatically. The particle swarm optimization (PSO) algorithm is commonly used for model parameter optimization in the field of bearing performance degradation assessment [8,22,23]. However, the traditional PSO algorithm suffers from slow convergence as well as local optimization problems, which affects the performance of the model. Hence, a modified PSO algorithm is suggested for the purpose of optimizing the LSTM model's parameters. The modified IPSO-LSTM module was applied to predict bearing performance degradation trends.

2. Methodology

2.1. LSTM

The mechanical degradation process, for example, on a rolling bearing, is a process of accumulation and continuous fault development [24]. Its degradation is determined by assessing its currently observable state as compared with its state in the recent past. The traditional neural network only uses the most recently documented state for its model, making it difficult to characterize deterioration and performance degradation over time. The LSTM is a type of RNN. An RNN is a neural network that handles sequential data and can be used to connect information from the recent past to the current task. However, as the distance between relevant information and the information taken from the past increases, the RNN loses its ability to learn and use distant details. Multiple control gates have been designed to replace the RNN in order to solve this problem. Thus, the LSTM network is constructed [16].

The LSTM solves the problem of gradient disappearance and explosion through the use of the aforesaid gates. In the LSTM structure, f_t, i_t, and o_t are three gates, which are designed to control the flow of information. f_t controls the information of memory cells

from time *t*-1 to time *t*. i_t controls the information input to the memory cells at time *t*, and o_t controls the information of memory cells at time *t* to the hidden state of h_t.

$$f_t = \sigma(w_{fc}C_{t-1} + w_{fh}h_{t-1} + w_{fx}x_t + b_f) \quad (1)$$

$$i_t = \sigma(w_{ic}C_{t-1} + w_{ih}h_{t-1} + w_{ix}x_t + b_i) \quad (2)$$

$$o_t = \sigma(w_{oc}C_{t-1} + w_{oh}h_{t-1} + w_{ox}x_t + b_o) \quad (3)$$

where w_{fc}, w_{ih}, and w_{oh} are the weight matrix between gate f_t and memory cell C_{t-1}. b_f is the bias of the gate f_t. Other weight matrices are derived from the following: C_t and C_{t-1} represent the values of memory cells at time *t* and time *t*-1. b_f, b_i, and b_o represent the bias. σ is the activation function. The hiding unit structure of the long and short-term memory network is shown in Figure 1.

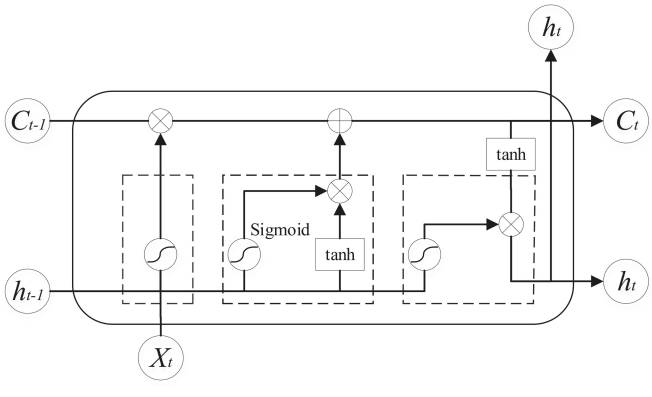


Figure 1. Structure of long short-term memory hidden unit.

The LSTM can predict degradation due to the time-varying characteristics of performance degradation and the advantages of LSTMs in modeling and forecasting time series. However, the structure of the LSTM model is complex. Some parameters need to be set synthetically, such as the time frame, the batch size, the number of hidden layer units, etc., which makes it difficult to meet the highly precise requirements for predicting time series degradation. Thus, it is necessary to find the optimal model parameters for each iteration in order to maintain strong predictive accuracy. This optimal model is realized through the use of a swarm intelligence algorithm, which auto-selects and optimizes the LSTM model's parameters to improve the prediction.

2.2. IPSO

A particle swarm optimization (PSO) algorithm is a population intelligent optimization algorithm used to simulate birds' foraging behavior. Kennedy and Eberhart first proposed it in 1995 [25]. A standard particle swarm optimization algorithm sets the particle swarm size as *m*, and each particle has an *n* dimension search region. $x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in})$ represents the search position of particle *I* in space. $v_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{in})$ is the velocity of the particle, *i*, which represents the moving distance of the particle in each position update. $p_i = (p_{i1}, p_{i2}, p_{i3}, \ldots, p_{in})$ records the search optimal bit value of the particle, *i*. $p_g = (p_{g1}, p_{g2}, p_{g3}, \ldots, p_{gn})$ is the optimal particle location in the current population. In a traditional PSO algorithm, the positions and velocities of particles are updated through Equations (4) and (5).

$$v_{ij}(t+1) = wv_{ij}(t) + c_1R_1(p_{ij}(t) - x_{ij}(t)) + c_2R_2(p_{gj}(t) - x_{ij}(t)) \quad (4)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (5)$$

where w is the inertia weight factor, and the range of w is (0, 1.4); c_1 and c_2 are learning factors; R_1 and R_2 are random numbers between 0 and 1; $v_{ij}(t+1)$ is the j dimension velocity component of the particle, i, in the t+1 iteration; $x_{ij}(t+1)$ is the j dimension position component of the particle, i, in the t+1 iteration; $p_{ij}(t)$ is the j dimension optimal position component of the particle, i, in the t iteration; $p_{gj}(t)$ is the j dimension position component of the optimal solution in the population in the t iteration; $1 \le i, g \le m$, $1 \le j \le n$.

However, the disadvantages of a PSO algorithm include low convergence accuracy and premature results. The diversity of the population decreases while the iteration times increase, the algorithm falls into the local optimal solution and the algorithm prematurely solves an incomplete problem. To solve this problem, the position updating model and parameter adjustment strategy of the particle swarm are modified.

1. Parameter adjustment strategy

The inertia weight, w, has a significant influence on the performance of particle swarm optimization. In the early stages, a strong search ability is needed to search for the best information quickly. And in the late stages, a fine selection is required to search for accuracy.

$$w^t = w_{\max} - t \times (w_{\max} - w_{\min})/t_{\max} \tag{6}$$

where $w_{\max}$ and $w_{\min}$ are the upper and lower limits of the preset inertia weight, and in general, $w_{\max} = 0.9$ and $w_{\min} = 0.4$, respectively; $t_{\max}$ is the maximum number of iterations.

In the traditional PSO algorithm, c_1 and c_2 are fixed values. The improved c_1 and c_2 can adaptively adjust learning factors and inertia weight. To find the optimal solution, independent and team learning abilities are adjusted in different search times.

$$c_1 = c_{\max} + (c_{\max} - c_{\min})(1 - (e^{-w} - 1)^2) \tag{7}$$

$$c_2 = c_{\max} - (c_{\max} - c_{\min})(1 - (e^{-w} - 1)^2) \tag{8}$$

2. Particle swarm position updating model [26]

In the early stages of the searching process, particles have a strong self-learning ability, and the search iterative step size should be set to large. With the search time increasing, the space range of solutions becomes smaller. To search for accurate solutions, the search iteration step size of particles should be reduced accordingly. Therefore, an adaptive adjustment factor, μ, is added to the particle position updating algorithm.

$$\mu = 1/(1 + e^{-t/t_{\max}}) + 1/2 \tag{9}$$

where t is the iterations.

The improved particle position update formula is then as follows.

2.3. IPSO-LSTM

As can be seen in Section 2.1, due to the advantage of processing time sequences, a two-layer LSTM is used as the backbone network for the high dimensional degradation feature extraction in this paper. The hidden state of each time step in the first layer is retained to serve as the input of the second layer, which only returns the hidden state of the last time step. To avoid model overfitting, a dropout regularization strategy is employed after each LSTM layer. Then, the learned representation features are fed into the fully connected layer to be mapped into a one-dimensional degeneration metric.

First, the hyper-parameters that need to be determined for the LSTM's backbone network include the number of hidden nodes in the first second layer. The hidden layers play a vital role in extracting high-dimensional features and internal laws. The model's performance is affected mainly by the number of hidden nodes. Too many nodes will increase the training time and may lead to overfitting. Too few will reduce the model's learning ability to the extent that the sparse adequate information extracted will not suffice

in solving the problem. Therefore, the model structure's complexity and predictive accuracy should be considered comprehensively in selecting the number of nodes when designing the network.

In addition, most neural networks are usually optimized by a gradient descent algorithm. The gradient descent is calculated as follows:

$$g = \frac{1}{m}\nabla_{\tilde{\theta}}\sum\nolimits_i L(f(x^{(i)};\tilde{\theta}), y^{(i)}) \tag{10}$$

where m is the batch size; $y^{(i)}$ is the target corresponding to m; θ is the updated parameter; f is the random target function with the parameter, θ.

As seen from Equation (10), increasing the batch size reduces the gradient and makes the gradient more accurate. This indicates that t stability of the convergence is enhanced by increasing batch size in the correct range.

As described above, it is clear that the three hyper-parameters, namely the number of hidden nodes in the first LSTM layer, the number of hidden nodes in the second LSTM layer, and batch size, are the key factors affecting the performance of the model. The specific representations and ranges are shown in Table 1. In this paper, the IPSO algorithm is used to optimize and automatically select the parameters of the LSTM model.

Table 1. Optimized Parameters.

Description	Notion
Number of nodes in the first LSTM layer	h_1
Number of nodes in the second LSTM layer	h_2
Batch size	Sm

h_1h_2 The flowchart of LSTM parameters optimized by IPSO is shown in Figure 2. The steps are as follows:

1. Initialize the parameters. Determine the population size range, iteration times, learning factors, location, and velocity;
2. Initialize the position and velocity of the particles. Generate the population particles $X_{i,0}(h_1, h_2, Sm)$ randomly. Where h_1 and h_2 denote the number of neurons in the first and second hidden layer, respectively, and Sm represents the batch size;
3. Determine the evaluative function of the particles. The particle $X_{i,0}$, in step 2 above, is assigned to the LSTM parameter. The data are partitioned into the training samples, validation samples, and test samples. The fitness value, fit, of individual X_i is defined as the target function, which is set as:

$$fit = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{11}$$

   where $\hat{y}_i$ is the predicted value; y_i is the actual observation;
4. Calculate the fitness value of each particle position, X_i. Individual extreme value and the population extreme value are determined according to the initial particle's fitness value, and each particle's best position is taken as its historical best position;
5. Update the velocity and position of the particle;
6. Determine whether the end condition of the iteration has been met. If it has, output the optimal parameter; Otherwise, go to step 4 to continue the iteration.

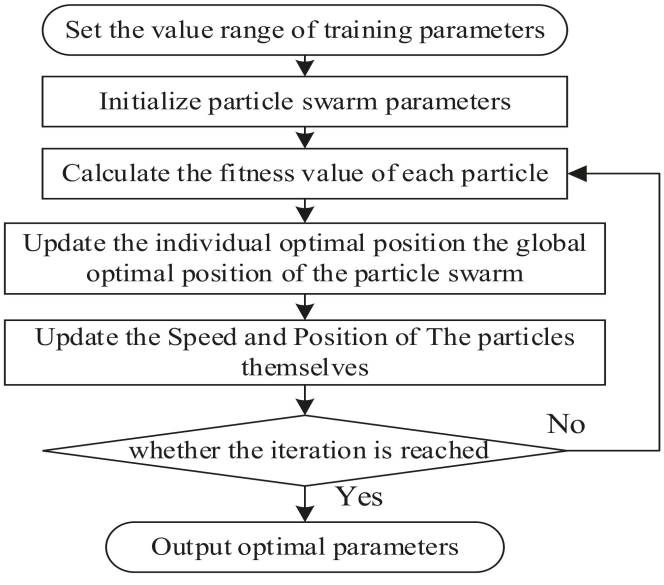


Figure 2. Parameter optimization flowchart.

3. Results IPSO-LSTM Method for Bearing Performance Degradation Prediction

In most practical industrial applications, the actual working conditions of mechanical equipment are complex and dynamic. Bearing vibration signals collected by sensors contain rich information. A single feature cannot fully describe the state of bearing vibration signals. The kernel joint approximate diagonalization of eigen-matrices (KJADE) is to map the observation data $X = \{x_1, x_2, \ldots, x_m\}$ to a high-dimensional feature space F through a nonlinear function Φ, and the mapped feature space is $F = \{\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_M)\}$. The inner product of two vectors in the feature space is calculated following the kernel function, and an $m \times m$ kernel matrix K is established as follows:

$$K_{ij} = k(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle \tag{12}$$

where x_i and x_j are the sample vectors. Therefore, the KJADE algorithm is employed to characterize the bearing degradation state.

The step flow chart of the method is shown in Figure 3. The operations are described as follows:

1. Original feature extraction. The full life vibration signal of bearing is analyzed in both the time and frequency domains to avoid the insufficiency of single feature evaluation ability. Eight features in time-domain and frequency-domain are extracted to form a high-dimensional feature vector, as shown in Table 2. $T1$–$T8$ are the mean value, root mean square (RMS), absolute average, skewness, waveform index, impulsion index, and kurtosis index, respectively. Among others are frequency domain features, where s_i is a spectrum for $i = 1, 2, \ldots, N$ (N is the number of spectrum lines) and f_i is the frequency value of the i-th spectrum line, indicating the degree of dispersion or concentration of the spectrum and the change of the dominant frequency band;

Table 2. Original features.

Time-domain	$T1 = \frac{1}{N}\sum_{i=1}^{N} x_i$, $T2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$, $T3 = \left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i\rvert}\right]^2$, $T4 = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert$, $T5 = \frac{1}{N}\sum_{i=1}^{N} x_i^3$, $T6 = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}}{F4}$, $T7 = \frac{\max(x)}{\frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert}$, $T8 = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i^4}{\left(\sqrt{\frac{1}{N}\sum_{j=1}^{N} x_j^2}\right)^4}$
Frequency-domain	$F1 = \frac{1}{N}\sum_{i=1}^{N} s_i$, $F2 = \frac{1}{N}\sum_{j=1}^{N}(s_j - \frac{1}{N}\sum_{i=1}^{N} s_i)^2$, $F3 = \frac{\sum_{i=1}^{N} f_i s_i}{\sum_{j=1}^{N} s_j}$, $F4 = \frac{\frac{1}{N}\sum_{j=1}^{N}(s_j - \frac{1}{N}\sum_{i=1}^{N} s_i)^3}{\left(\sqrt{F10}\right)^3}$, $F5 = \sqrt{\frac{\sum_{i=1}^{N} f_i^2 s_i}{\sum_{j=1}^{N} s_j}}$, $F6 = \sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i (f_i - F12)^2}$, $F7 = \sqrt{\frac{\sum_{i=1}^{N} f_i^4 s_i}{\sum_{j=1}^{N} f_j^2 s_j}}$, $F8 = \frac{\sum_{i=1}^{N} f_i^2 s_i}{\sqrt{\sum_{j=1}^{N} s_j \sum_{k=1}^{N} f_k^4 s_k}}$

2. KJADE features fusion. Considering the nonlinear characteristics of bearing vibration, the redundancy of the original multi-domain degradation characteristics, and some features that are not sensitive to the bearing degradation state, it is necessary to fuse multi-domain features. Therefore, the KJADE algorithm based on kernel function is employed to extract more effective, but low-dimensional, degradation characteristic indexes [27]. KJADE has better nonlinear processing capabilities for bearing vibration signals. It maps the observation data to a high-dimensional feature space through a nonlinear function. Then the JADE can be used in this feature space to change the nonlinear separable problem into a linear one;
3. Degradation assessment index calculation. The vibration signal collected at the beginning of the bearing operation is taken as the health state, corresponding to the health sample after dimension reduction. The subsequent signal is selected as the monitoring sample. To quantify the dispersion between monitoring and health samples and the aggregation between different classes, the evaluation factor, SS, composed of between- and within-class scatter matrix, is used as the performance degradation index [12];
4. IPSO-LSTM model construction. The number of hidden layer nodes and the batch size of the LSTM neural network are taken as optimization objects. The LSTM is constructed according to the corresponding parameters of each particle. The IPSO algorithm is used to acquire the optimal hyper-parameter set for each iteration automatically;
5. Predicting the performance degradation. The LSTM model is constructed with the optimal value of hyper-parameters, and the bearing data is used as input for training and prediction.

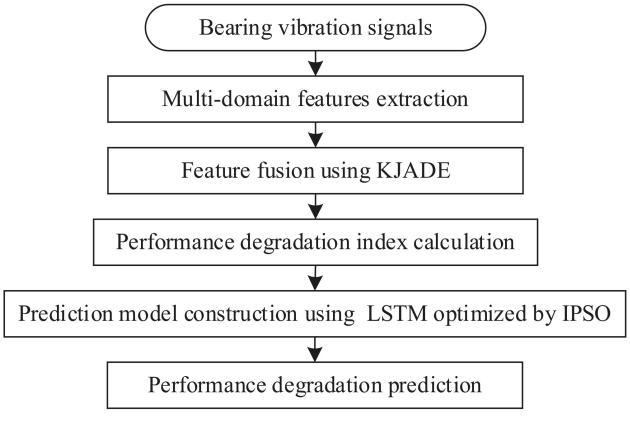


Figure 3. Performance degradation prediction by IPSO.

4. Case Analysis

4.1. Case 1

The Intelligent Maintenance System (IMS) Center of the University of Cincinnati's full-life vibration signals of bearings are used to confirm the proposed method [28]. The experimental platform is shown in Figure 4.

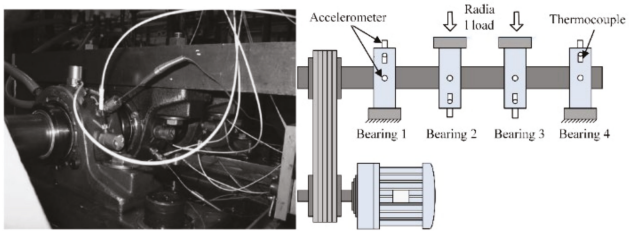


Figure 4. Experimental setup.

The bearing type is ZA-2115, and the experimental conditions were as follows: output speed was 2000 rpm, the radial load was 6000 lbs, and the sampling frequency was 20,480 Hz. A total of 984 sets of vibration signal data were recorded. The whole experiment was completed in three groups. By the end of the experiment, an inner fault in bearing 3 and a rolling fault in bearing 4 were observed in the first group. An outer fault in bearing 1 in the second group and an outer fault in bearing 3 in the third group were also observed. Among them, the rolling fault and inner fault in the first group, along with the outer fault in the second group, were selected as objects for analysis. The corresponding vibration data of life is shown in Figure 5.

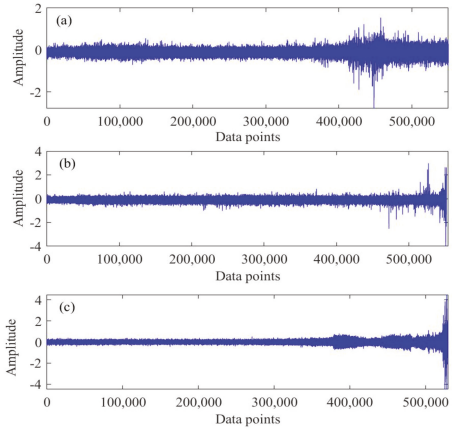


Figure 5. Exemplary diagram of bearing vibration data: (**a**) rolling fault; (**b**) inner fault; (**c**) outer fault.

Based on the method in Section 2, the IPSO algorithm is used to optimize the LSTM model's predictive parameters. The initial parameters of the IPSO are as follows: the number of particles is 10, the dimension of particle swarm is 3, the maximum velocity of the particle is 1, and the maximum iteration number is 50. The range of particle locations, namely the number of hidden layer nodes, is set to (100, 300), and the batch size is (30, 200). The upper and lower limits of the inertia weight are w_max = 0.9 and w_min = 0.5, while the upper and lower limits of the initial learning factors c_max and c_min are 2 and 1, respectively. These are the optimal parameters obtained by comparative experiments. In this study, the first 60% of the performance data is used as the training set, and 20% of the

rest is saved as a validation set. Besides this, the model is optimized by an Adam algorithm, and the root mean square error (RMSE) is applied as the target criteria.

To demonstrate the superiority of the proposed method, the performance of conventional LSTMs and PSO-LSTMs have been compared. The resulting real degradation trends, which can be expressed as a degradation index, are obtained via feature fusion using the KJADE algorithm. Additionally, the comparison results of the degradation trends predicted by each model are shown in Figures 6–8, where the y-axis is the degradation index. In addition, RMSE is used as an additional metric to measure the performance of the model, with the results shown in Table 3. The RMSE calculation is shown in Equation (13).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{13}$$

where $\hat{y}_i$ is the predicted value; y_i is the actual observation; n is the total number of samples in the faulty bearing.

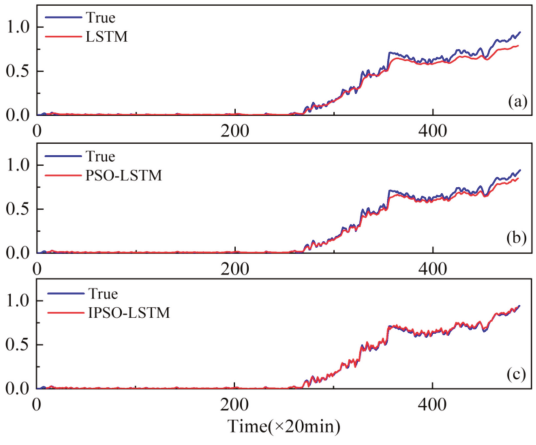


Figure 6. Performance degradation predictions for outer bearings for the: (**a**) LSTM; (**b**) PSO-LSTM; (**c**) IPSO-LSTM.

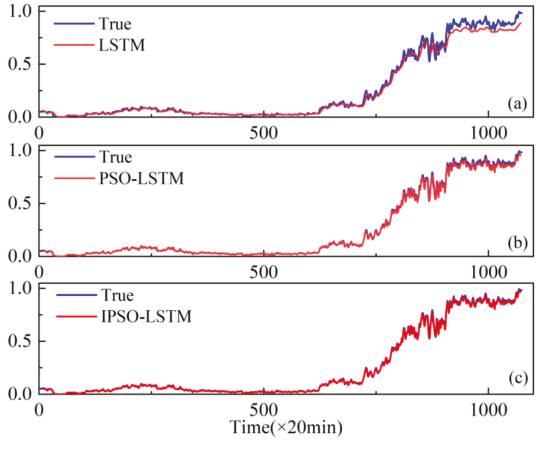


Figure 7. Performance degradation predictions of roller bearings for the: (**a**) LSTM; (**b**) PSO-LSTM; (**c**) IPSO-LSTM.

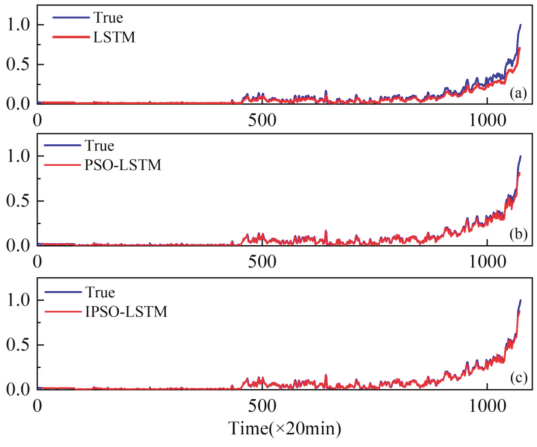


Figure 8. Performance degradation predictions of inner bearings for the: (**a**) LSTM; (**b**) PSO-LSTM; (**c**) IPSO-LSTM.

Table 3. The RMSE of LSTM with different optimization methods.

RMSE	Outer	Roller	Inner
LSTM	0.042	0.039	0.042
PSO-LSTM	0.025	0.013	0.018
IPSO-LSTM	0.012	0.011	0.013

From Figures 6–8, it can be seen that our proposed IPSO-LSTM method tracks the degenerate states significantly better than the other two methods in all three failure modes, especially the LSTM method without the hyper-parameter optimization process. In terms of quantitative metrics, the RMSE results in Table 3 also illustrate the superiority of the proposed method.

The above results show that the IPSO algorithm is effective in optimizing the hyper-parameters of the LSTM based network, which can automatically and accurately search for the optimal parameters. To further illustrate the advantages of the IPSO algorithm in optimizing speed and avoiding local extremum, we visualize the parameter search processes, which are shown in Figure 9.

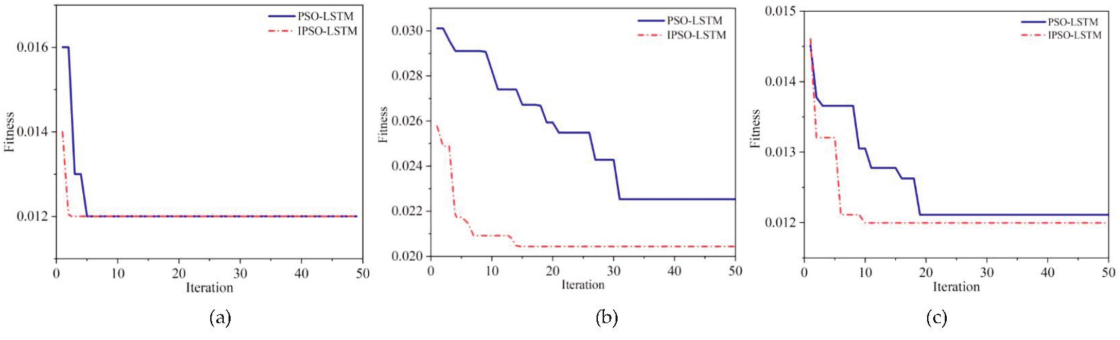


Figure 9. Optimization iteration results for the: (**a**) outer bearing; (**b**) roller bearing; (**c**) inner bearing.

Overall, the convergence speed and fitness of the IPSO algorithm are better than the traditional PSO algorithm. Specifically, as Figure 9b,c demonstrate, IPSO has good

optimization ability and can quickly find the optimal global point. Compared with the PSO, the IPSO algorithm has a faster convergence speed. Figure 9a shows that although the final fitness error is the same, the IPSO algorithm converge is faster.

Furthermore, extreme learning machines (ELM) and support vector regression (SVR), which have been widely used with good performance degradation prediction [29,30], are compared with the proposed IPSO-LSTM for effectiveness. The comparison results are shown in Figures 10–12.

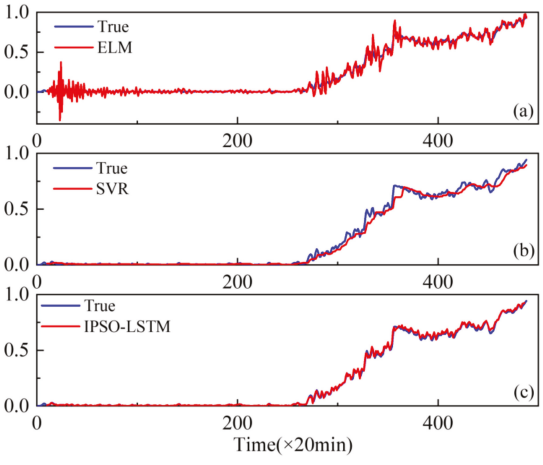


Figure 10. Performance degradation predictions of the outer bearings for the: (**a**) ELM; (**b**) SVR; (**c**) IPSO-LSTM.

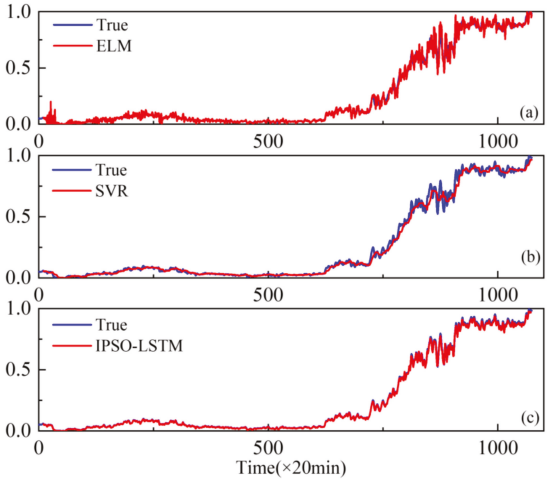


Figure 11. Performance degradation predictions of roller bearings for the: (**a**) ELM; (**b**) SVR; (**c**) IPSO-LSTM.

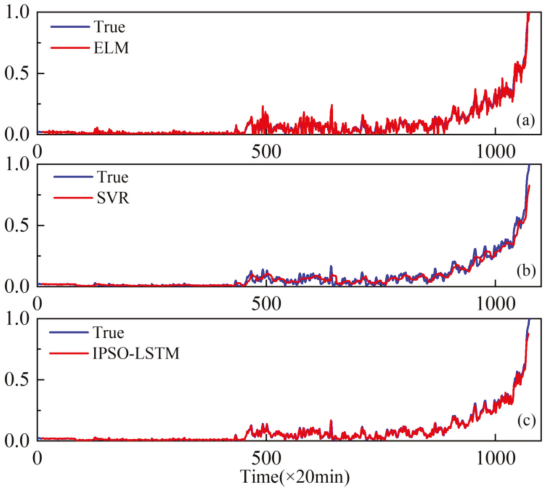


Figure 12. Performance degradation predictions of the inner bearings for the: (**a**) ELM; (**b**) SVR; (**c**) IPSO-LSTM.

The results show that the prediction results of the IPSO-LSTM method are more in line with the original curve, with greater predictive accuracy. This is demonstrated in the RMSE values in Table 4. Predictive errors in the proposed method are minimal, which verifies the effectiveness of the proposed IPSO-LSTM method.

Table 4. The RMSE of different methods.

RMSE	Outer	Roller	Inner
ELM	0.055	0.031	0.029
SVR	0.031	0.027	0.029
IPSO-LSTM	0.012	0.011	0.013

4.2. Case 2

The lab experiments used four HRB6305 bearings. They were fixed on the same shaft and connected with the motor. A radial load of 750 kg was applied to all bearings to accelerate the bearing damage process, and the bearing speed was 3000 rpm. Full-life vibration signals were obtained by the NI PXI acquisition system. The vibration signals acquisition frequency was 20 kHz, the data were collected every 5 min. The experimental platform is shown in Figure 13.

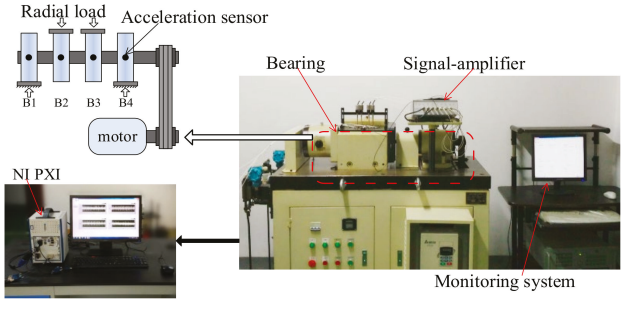


Figure 13. Experimental setup.

The fault in the rolling element is taken as the experimental object. Figure 14 shows the full-life original vibration signal of the rolling element. The mixed-domain features are extracted from the bearing data. KJADE is used for feature fusion to acquire an optimal feature parameter set, and the SS is calculated from fusion features to obtain the degradation index. The proposed method is used to predict the performance degradation and compared with the LSTM and PSO-LSTM methods. The prediction curve is shown in Figure 15.

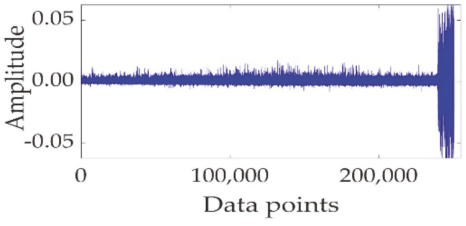


Figure 14. The full-life original vibration signal.

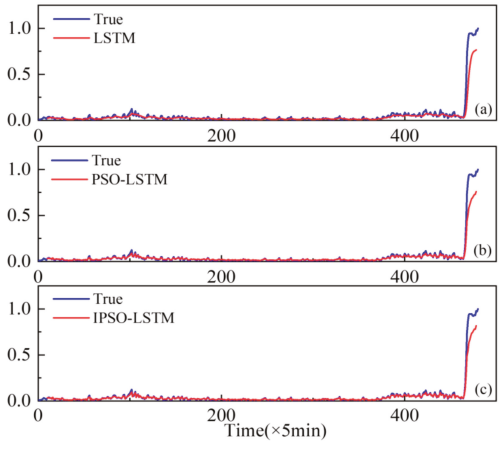


Figure 15. Performance degradation predictions of the roller bearings for the: (**a**) LSTM; (**b**) PSO-LSTM; (**c**) IPSO-LSTM.

The results demonstrate that the predictive accuracy of the proposed method is greater than that of the other two methods. The RMSE results of LSTM, PSO-LSTM, and IPSO-LSTM are shown in Table 5. The iteration results of IPSO and PSO optimization are shown in Figure 16. It demonstrates that the IPSO algorithm converges earlier and is less likely to succumb to the local minimum problem, which is an advantage over the performance of the PSO.

Table 5. The RMSE of LSTM with different optimization methods.

	LSTM	**PSO-LSTM**	**IPSO-LSTM**
RMSE	0.065	0.054	0.048

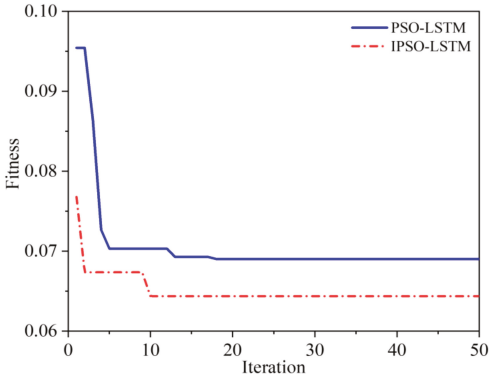


Figure 16. Optimization iteration results of roller bearing.

Similar to case 1, extreme learning machines (ELM) and support vector regression (SVR) are compared with the proposed method.

The results of the comparison are shown in Figure 17 and Table 6. It can be seen that the proposed method is more effective than the other two methods in predicting the degradation trend of bearings. The RMSE values also reflect that the proposed IPSO-LSMT's predictive accuracy is higher than the ELM and SVR methods.

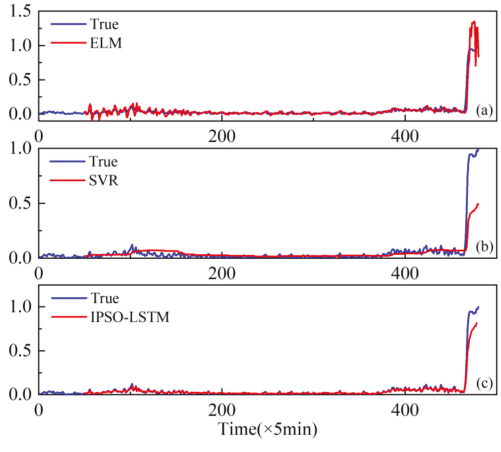


Figure 17. Performance degradation predictions of the roller bearing for the: (**a**) ELM; (**b**) SVR; (**c**) IPSO-LSTM.

Table 6. The RMSE of different methods.

	ELM	SVR	IPSO-LSTM
RMSE	0.073	0.101	0.048

5. Conclusions

This paper proposes a method based on an improved PSO optimized LSTM (IPSO-LSTM) to analyse bearing performance degradation. The proposed method can effectively resolve the problem of online parameter selection and the low predictive accuracy of the LSTM method. The KJADE method is used to fuse the bearing vibration signal to form an effective feature vector, and SS is calculated to acquire a performance degradation index. Then, the improved PSO algorithm is used to optimize the LSTM parameters to

obtain an optimal performance degradation prediction model. In this study, the proposed method is compared with the LSTM, PSO-LSTM, ELM, and SVR through lab experiments. The experiments' results have verified the effectiveness and superiority of the proposed method over others. This method has good prospective applications in predicting bearing performance degradation, and it can also be tailored and applied to other mechanical systems for online health and prognosis management.

Author Contributions: Conceptualization, Y.H. and Y.L.; methodology, Y.Y. and R.W.; software, R.W.; validation, X.L., Z.H. and Y.H.; formal analysis, Y.L.; investigation, C.H. and R.W.; resources, Y.Y.; data curation, Z.H.; writing—original draft preparation, Y.H. and R.W.; writing—review and editing, H.L.; visualization, H.L.; supervision, Y.L.; project administration, Y.Y.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China [52075001, 52105082, 52105040, 52075002]; and the Key Basic Research Project [MKF20210008].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated, or that appeared in this study, are available upon request by contact with the corresponding author. Furthermore, the models and codes used during the study cannot be shared at this time as the data also forms part of an ongoing study.

Acknowledgments: The authors are grateful to the editors and anonymous reviewers for their helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rai, A.; Upadhyay, S.H. The use of MD-CUMSUM and NARX neural network for anticipating the remaining useful life of bearings. *Measurement* **2017**, *111*, 397–410. [CrossRef]
2. Nohal, L.; Vaculka, M.; Iop. Experimental and computational evaluation of rolling bearing steel durability. In Proceedings of the 4th International Conference Recent Trends in Structural Materials (COMAT), Pilsen, Czech Republic, 9–11 November 2017.
3. Saravanan, N.; Ramachandran, K.I. Incipient gear box fault diagnosis using discrete wavelet transform (DWT) for feature extraction and classification using artificial neural network (ANN). *Expert Syst. Appl.* **2010**, *37*, 4168–4181. [CrossRef]
4. Zhang, J.; Sun, Y.; Guo, L.; Gao, H.; Hong, X.; Song, H. A new bearing fault diagnosis method based on modified convolutional neural networks. *Chin. J. Aeronaut.* **2020**, *33*, 439–447. [CrossRef]
5. Zhu, J.; Chen, N.; Peng, W. Estimation of Bearing Remaining Useful Life Based on Multiscale Convolutional Neural Network. *Ieee Trans. Ind. Electron.* **2019**, *66*, 3208–3216. [CrossRef]
6. Tian, Z. An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. *J. Intell. Manuf.* **2012**, *23*, 227–237. [CrossRef]
7. Zheng, J.; Pan, H.; Cheng, J. Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines. *Mech. Syst. Signal Processing* **2017**, *85*, 746–759. [CrossRef]
8. Zhang, B.; Zhang, S.; Li, W. Bearing performance degradation assessment using long short-term memory recurrent network. *Comput. Ind.* **2019**, *106*, 14–29. [CrossRef]
9. Li, X.; Jiang, H.; Wang, R.; Niu, M. Rolling bearing fault diagnosis using optimal ensemble deep transfer network. *Knowl.-Based Syst.* **2021**, *213*, 106695. [CrossRef]
10. Hu, M.; Wang, G.; Ma, K.; Cao, Z.; Yang, S. Bearing performance degradation assessment based on optimized EWT and CNN. *Measurement* **2021**, *172*, 108868. [CrossRef]
11. Chen, C.; Zhang, B.; Vachtsevanos, G. Prediction of Machine Health Condition Using Neuro-Fuzzy and Bayesian Algorithms. *Ieee Trans. Instrum. Meas.* **2012**, *61*, 297–306. [CrossRef]
12. Ren, L.; Sun, Y.Q.; Wang, H.; Zhang, L. Prediction of Bearing Remaining Useful Life With Deep Convolution Neural Network. *IEEE Access* **2018**, *6*, 13041–13049. [CrossRef]
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
14. Malhi, A.; Yan, R.Q.; Gao, R.X. Prognosis of Defect Propagation Based on Recurrent Neural Networks. *IEEE Trans. Instrum. Meas.* **2011**, *60*, 703–711. [CrossRef]
15. Zhao, R.; Wang, D.Z.; Yan, R.Q.; Mao, K.Z.; Shen, F.; Wang, J.J. Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks. *IEEE Trans. Ind. Electron.* **2018**, *65*, 1539–1548. [CrossRef]
16. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

17. Ma, M.; Mao, Z. Deep-Convolution-Based LSTM Network for Remaining Useful Life Prediction. *IEEE Trans. Ind. Inform.* **2021**, *17*, 1658–1667. [CrossRef]
18. Liu, G.; Zhao, J.; Zhang, X. Bearing degradation trend prediction under different operational conditions based on CNN-LSTM. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *612*, 032042. [CrossRef]
19. Elsheikh, A.; Yacout, S.; Ouali, M.S. Bidirectional handshaking LSTM for remaining useful life prediction. *Neurocomputing* **2019**, *323*, 148–156. [CrossRef]
20. Tang, G.; Zhou, Y.; Wang, H.; Li, G. Prediction of bearing performance degradation with bottleneck feature based on LSTM network. In Proceedings of the 2018 IEEE International Instrumentation and Measurement Technology Conference, I2MTC 2018, Houston, TX, USA,, 14–17 May 2018; pp. 1–6.
21. Zhao, Z.; Wu, J.; Li, T.; Sun, C.; Yan, R.; Chen, X. Challenges and Opportunities of AI-Enabled Monitoring, Diagnosis & Prognosis: A Review. *Chin. J. Mech. Eng.* **2021**, *34*, 56. [CrossRef]
22. Ding, N.; Li, H.; Yin, Z.; Zhong, N.; Zhang, L. Journal bearing seizure degradation assessment and remaining useful life prediction based on long short-term memory neural network. *Measurement* **2020**, *166*, 108215. [CrossRef]
23. Rathore, M.S.; Harsha, S.P. Prognostics Analysis of Rolling Bearing Based on Bi-Directional LSTM and Attention Mechanism. *J. Fail. Anal. Prev.* **2022**, 1–20. [CrossRef]
24. Wang, B.; Lei, Y.; Yan, T.; Li, N.; Guo, L. Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery. *Neurocomputing* **2020**, *379*, 117–129. [CrossRef]
25. Marini, F.; Walczak, B. Particle swarm optimization (PSO). A tutorial. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 153–165. [CrossRef]
26. Kang, Y.; Jiang, C.; Qin, Y.; Ye, C. Robot Path Planning and Experiment with an Improved PSO Algorith. *Robot* **2020**, *42*, 8. [CrossRef]
27. Liu, Y.B.; He, B.; Liu, F.; Lu, S.L.; Zhao, Y.L. Feature fusion using kernel joint approximate diagonalization of eigen-matrices for rolling bearing fault identification. *J. Sound Vib.* **2016**, *385*, 389–401. [CrossRef]
28. Gousseau, W.; Antoni, J.; Girardin, F.; Griffaton, J. Analysis of the Rolling Element Bearing data set of the Center for Intelligent Maintenance Systems of the University of Cincinnati. In Proceedings of the CM 2016, Charenton, France, 10 October 2016.
29. Fang, L.; Yongbin, L.; Fenglin, C.; Bing, H. Residual life prediction for ball bearings based on joint approximate diagonalization of eigen matrices and extreme learning machine. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2017**, *231*, 1699–1711.
30. Liu, F.; Li, L.; Liu, Y.; Cao, Z.; Lu, S. HKF-SVR Optimized by Krill Herd Algorithm for Coaxial Bearings Performance Degradation Prediction. *Sensors* **2020**, *20*, 660. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Sensors Editorial Office
E-mail: sensors@mdpi.com
www.mdpi.com/journal/sensors

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34

www.mdpi.com

ISBN 978-3-0365-7333-5